On Logical System Access Control and the Associated User and Network Management in Future Heterogeneous 4G Wireless Systems

Artur Hecker

École Doctorale
d'Informatique,
Télécommunications
et Électronique de Paris

Thèse

présentée pour obtenir le grade de Docteur
de l'École Nationale Supérieure des Télécommunications

Spécialité : **Informatique et Réseaux**

ARTUR HECKER

Contrôle d'accès et gestion des réseaux 4G hétérogènes

Soutenue le 16 mars 2005 devant le jury composé de

Guy Pujolle	Président
Adam Wolisz	Rapporteurs
Jochen Schiller	
Djamal Zeghlache	
Ahmed Serhrouchni	Examinateur
Marc Challand	Invités
Pierre Lévis	
Houda Labiod	Directeur de thèse

On Logical System Access Control and the Associated User and Network Management in Future Heterogeneous 4G Wireless Systems

Applying new paradigms to overcome heterogeneity

PhD Thesis – Thèse de Doctorat
at the Département Informatique et Réseaux,
l'Ecole Nationale Supérieure des Télécommunications,
Paris, France

by

Artur Hecker

February 16, 2005

After years of a continuous observation I finally understood that my unsettled life is revolving around one single fixed point. Much like the Sun for the Earth, this center of gravity is my personal source of energy and provides me a stable trajectory in everything I undertake.

To my loving family.

Acknowledgements

A lot of people have directly or indirectly contributed to this work. This section is a humble try to express my gratitude and appreciation.

Firstly, I would like to thank my thesis supervisor, Dr. Houda Labiod (ENST), who initiated this work, organized the framework for it and guided me through to its successful end. I especially thank her for teaching me the methodology of the scientific work. Warmest thanks for all the time spent on me and the motivation pushes given to me during the hard times. I guess I needed it.

Secondly, this work would not have been possible without the Département Informatique et Réseaux (INFRES, Department of Computer Science and Networks) at the ENST Paris and the Ecole Doctorale d'Informatique, Télécommunications et Electronique de Paris (EDITE) who provided me place, funding and support during the three years. Thank you very much.

Next, I would like to thank all the jury members for the time and energy which they have spent on my PhD defense. I feel deeply honored by the attention brought to my modest contributions by these experts and would like to thank them for the interesting comments, questions and discussions before, during and even after the defense. Thanks to Marc Challand (EADS) and Pierre Lévis (France Télécom) for their time and patience. My particular thanks to Pr. Guy Pujolle (LIP6/Paris VI), Dr. Ahmed Serhrouchni (ENST) and Pr. Djamal Zeghlache (INT Evry) who accompanied my ongoing research through the common work in various projects. I would like to express thanks to Prof. Dr.-Ing. Jochen Schiller (FU Berlin) for his constructive review. Interestingly, I first came in touch with mobile communications in his class in the late 90s and I am convinced that his inciting teaching is the basis for this work. I also feel very grateful for the encouraging review of

Prof. Dr.-Ing. Adam Wolisz (TU Berlin), for his coming in spite of the short schedule and, particularly, the fine discussion about the profession which we had after the defense.

For fruitful discussions and often surprising insights in different matters I would also like to thank Michel Riguidel (Head of INFRES, ENST), Dr. Gwendal LeGrand (ENST) and Dr. Jean Leneutre (ENST).

A PhD thesis is a long way to go and obstacles are rather normal than surprising. Very special thanks thus go to my closest friends in Paris who continuously encouraged, comforted and motivated me during this three-years-run through all the hurdles. I would like to express my deep appreciation to the fine gentlemen Sassan Rostambeik (ENST), Franck Springinsfeld (WaveStorm) and Emilio Calvanese Strinati (ENST). Their help included direct problem solving, professional advice, private life organization and personal concerns. Additionally, I would like to thank Franck for his active participation in two research projects and the development of big parts of the EAP-SIG prototype.

Both scientific and personal support was highly appreciated from Erik-Oliver Blaß (University of Karlsruhe, Germany). As my most intimate friend back in Karlsruhe, he always found time for me and my preoccupations.

Last but definitely not least, I would like to thank the following:

- Vincent Guyot (LIP6) for the willing help in the daily life,
- Ann Törnkvist for proof-reading,
- all the colleagues and friends at WaveStorm, Paris,
- complete staff at the INFRES department, ENST.

And of course, I owe a lot to my family. Danke, dass Ihr an mich immer geglaubt habt. Ich liebe Euch.

Artur Hecker
Paris, Summer 2005

Résumé en français

Les futurs réseaux des télécommunications mobiles dépasseront largement le transport de la voix dans les réseaux mobiles classiques (2e Génération) et le transport des données dans les réseaux mobiles émergeant (3e Génération). Nous poursuivons l'idée de la prochaine génération des systèmes de télécommunications mobiles capable de fournir globalement un accès adaptatif aux services, prenant en compte le contexte actuel de l'utilisateur.

Nous identifions et discutons les problèmes liés aux technologies sans fil et mobiles actuelles et argumentons pour une vision de la 4e Génération en tant qu'un système intégrant d'une manière opportuniste des technologies hétérogènes et se concentrant sur l'utilisateur et ses besoins. Cette vision permet ainsi de combiner les avantages économiques et technologiques de différents standards de transmission.

En contrepartie, plusieurs nouveaux défis technologiques émergent de cette vision 4G spécifique. En particulier, nous identifions le contrôle d'accès aux services et la gestion des utilisateurs et de l'infrastructure en tant que défis majeurs et points critiques. En effet, de nouvelles solutions semblent nécessaires pour protéger les infrastructures déployées et, en même temps, pour offrir l'accès convivial aux services, disponible partout, à chaque instant et de manière continue. Actuellement, le contrôle d'accès utilisé est typiquement lié à la technologie utilisée, et le modèle économique existe sous forme d'une souscription fixe. Etant mis en place par l'opérateur, les deux sont orientés vers ses besoins et, du point de vue de l'utilisateur, limitent l'accès aux services à la technologie de cet opérateur et souvent à son infrastructure. Dans notre vision 4G, l'accès aux services est limité uniquement par le profil de l'utilisateur qui seul représente la situation contractuelle. En particulier, l'accès aux services doit être indépendant de l'opérateur et

de l'infrastructure. Il doit être adaptatif au niveau des mécanismes pour pouvoir prendre en compte le contexte actuel de l'utilisateur et changer l'association en cas de besoin. Par ailleurs, il doit répondre aux besoins de l'opérateur permettant un contrôle rapide, simple et garantissant le passage à l'échelle.

Dans le contexte du problème général de l'accès aux services logique, nous présentons multiples problèmes spécifiques liés aux systèmes hétérogènes. Nous discutons les solutions proposées et démontrons les limites technologiques du contrôle d'accès au niveau des couches basses et les problèmes de sécurité critiques propres au contrôle d'accès au niveau des couches hautes. En appliquant de nouveaux paradigmes, nous proposons une nouvelle solution basée sur les couches basses et la virtualisation des mécanismes du contrôle d'accès ramenant ceux-ci sur les couches hautes. Nous montrons comment cette solution peut être réalisée dans notre vision 4G. Nous présentons une implémentation illustrant cette approche au moyen d'un exemple d'un réseau campus basé sur un standard sans fil très répandu.

Concernant la gestion des utilisateurs et du réseau, nous étudions d'abord les possibilités du contrôle d'accès centralisé et discutons les problématiques et les améliorations possibles dans le cadre de la 4G. Nous discutons le besoin de la séparation de l'interface d'utilisateur de l'organisation interne des réseaux des opérateurs et présentons deux approches différentes à une telle organisation dans le cadre 4G. Nous présentons ensuite une nouvelle proposition appliquant les paradigmes égal-à-égal (P2P) dans le plan de contrôle. Nous montrons les avantages et les problématiques ouvertes avec cette approche. Nous discutons comment cette approche peut être implémentée dans un réseau local sans fil. Nous discutons également la co-existence d'une telle approche distribuée avec les approches centralisées classiques du contrôle d'accès dans les scénarii du nomadisme.

Finalement, nous identifions les thèmes ouverts et présentons les perspectives pour des recherches futures.

I.1. Introduction

Le sujet de cette thèse se situe au cœur de la problématique posée actuellement dans le contexte de passage des réseaux des télécommunications mobiles actuels à la prochaine génération. Jusqu'ici le changement des générations s'effectuait principalement sous la forme du changement de la technologie du lien entre l'utilisateur et le réseau. Toutefois, l'intérêt en technologies de transmission sans fil est actuellement énorme et le développement rapide dans le domaine des télécommunications numériques suggère une ouverture des systèmes des télécommunications mobiles vers une vision plus générale permettant l'intégration de toute nouvelle technologie dans un nouveau standard englobant ces méthodes d'accès. En plus de la considération des problématiques du lien 4G telles que par exemple débit, sécurité et qualité du lien, une telle vision nécessite la considération de l'architecture du réseau cœur et la réalisation dans cette architecture de l'accès aux mêmes services par différentes technologies.

Cette thèse a pour but d'étudier les opportunités et les problématiques imposées par la vision de la 4^{e} Génération des réseaux mobiles en tant que système intégrant d'une manière opportuniste les différentes technologies de transmission disponibles pour permettre l'accès aux services dans les réseaux des opérateurs liés par un réseau cœur utilisant la technologie Internet.

Dans ce résumé, nous éclaircissons les problématiques étudiées dans le vaste domaine de 4G et présentons ensuite nos propositions principales développées durant la thèse.

I.2. Problématiques étudiées

En adoptant une vision de 4G en tant que système hétérogène complexe, nous avons effectué une analyse du système. Les résultats de cette analyse nous ont amenés à la conclusion que le système envisagé possède trois interfaces principalement indépendantes. De plus, la préservation de cette indépendance dans le processus de la conception du système finale est essentielle pour la simplification du système résultant.

Notre analyse a identifié le réseau de l'opérateur comme élément architectural central. Dans le système 4G les instances de cette entité interférent avec les utilisateurs par l'interface utilisateur-opérateur et avec d'autres instances du même type par l'interface inter-opérateur. Outre cela, pour remplir son but cette entité possède des interfaces internes. Pour faciliter l'investissement et assurer la longévité d'un tel réseau, il est indispensable de préserver la flexibilité des opérateurs, c'est-à-dire concevoir le système final en préservant la vue sur cette entité comme sur une boîte noire implémentant les interfaces standardisées. Cette approche, que nous appelons la virtualisation, est comparable aux approches de programmation orientée objet. Dans ce contexte la virtualisation sépare logiquement les interfaces et les implémentations.

Avec cette conception, il est naturellement assuré que tout utilisateur peut potentiellement utiliser les services dans tout réseau de l'opérateur. Nous étudions les problématiques 4G en les associant à une des trois interfaces.

Nous identifions notamment l'accès logique aux services dans ce système hétérogène comme une problématique majeure. Nous présumons que l'accès physique est fourni par le standard de transmission sans fil. En contrepartie, l'accès logique aux services inclut des problématiques comme la découverte du réseau, la découverte des services, le contrôle d'accès en tant que fonction de sécurité, la configuration dynamique du terminal, etc. Les méthodes spécifiées dans le standard de transmission ne pouvant pas être utilisées à cause de leur hétérogénéité, de nouvelles méthodes doivent être développés tout en prenant en compte les limites imposées, par exemple par les technologies hétérogènes au niveau de sécurité du système, par les ressources du terminal au niveau de faisabilité, etc. Dans notre approche, cette problématique est principalement la problématique de l'interface utilisateur-opérateur. Toutefois, nous définissons l'utilisateur comme l'identifiant abstrait d'une personne physique faisant objet d'un contrat de service entre cette dernière et un opérateur. Cette définition de l'utilisateur implique que l'accès aux services peut provoquer des échanges entre deux opérateurs ce qui fait que le contrôle d'accès concerne également l'interface inter-opérateur, principalement sur le plan contrôle. Par conséquent, des spécifications précises sont nécessaires sur ces deux interfaces pour résoudre la problématique d'accès logique.

Ensuite, nous étudions les interfaces internes des opérateurs. Même si leur implémentation ne doit pas être spécifiée, il est essentiel de développer des architectures de références pour plusieurs scénarii pour montrer la faisabilité du système complet dans un premier temps et pour simplifier le déploiement plus tard. Le caractère d'un opérateur dans ce système 4G varie entre les petits opérateurs locaux jusqu'aux opérateurs nationaux. La principale problématique dans ce cadre semble être le besoin de réutilisation des infrastructures déployées pour assurer la flexibilité de l'opérateur et la sécurité de son investissement ; cela concerne les infrastructures physiques et logiques mais aussi la conception d'un plan de contrôle commun capable de gérer les équipements hétérogènes dans les réseaux de toutes tailles.

I.3. Contributions

Notre participation contient plusieurs nouvelles propositions pour des problématiques pointues dont les principales sont nommées ci-dessus. Nous traitons d'abord les problématiques liées au groupe d'interfaces à spécifier et passons ensuite aux problématiques internes du réseau de l'opérateur.

I.3.1. Accès aux services dans le cadre 4G

Optimisation du nomadisme

Dans une première partie nous avons proposé des améliorations aux échanges typiques entre les réseaux de deux opérateurs dans le cas de roaming ou nomadisme. Nous avons identifié l'interdépendance entre les échanges sur l'interface utilisateur-opérateur et l'interface inter-opérateur, notamment entre la réactivité du système aux changements du profil d'un utilisateur et le délai d'authentification dans le cas de roaming. Pour résoudre ce problème, nous avons proposé une séparation plus stricte entre l'autorisation d'accès et l'authentification de l'utilisateur. Sur la base des protocoles standards, nous avons développé un système éliminant cette interdépendance non souhaitable. Nous avons ensuite développé une plateforme de test entre deux écoles du GET que nous avons pu utiliser pour mesurer l'avantage de notre approche. Tout en préservant la réactivité immédiate aux changements du profil, notre approche procure jusqu'à 80% moins de messages échangés entre les deux opérateurs et diminue ainsi efficacement le délai d'accès.

Virtualisation de la signalisation

Dans la deuxième partie, nous avons étudié le dilemme suivant : la découverte des informations sur les réseaux disponibles est nécessaire pour pouvoir choisir le réseau ; toutefois, en présence de contrôle d'accès sur les couches basses exigé par la protection de l'infrastructure déployée une telle découverte n'est possible qu'après l'accès au réseau. Ce problème est très complexe : en effet, le contrôle d'accès sur les couches basses dans ce système hétérogène représente déjà un problème en soi car les couches basses changent en fonction de la technologie utilisée et il devient difficile d'assurer le même niveau de sécurité tel qu'il est exigé par le profil de l'utilisateur. Le besoin d'avoir des informations supplémentaires sur un réseau afin de pouvoir faire le choix du réseau à utiliser aggrave ce problème car une communication préalable à l'accès même devient nécessaire : la signalisation. Une telle communication doit être universelle pour pouvoir être utilisée dans tout réseau rencontré et devrait fournir par exemple des informations sur les services utilisables par l'utilisateur et leurs paramètres. La signalisation doit également permettre la configuration du terminal pour l'utilisation de tels services. Par conséquent, la signalisation se base toujours sur les couches hautes. Cela constitue le cœur du problème : le contrôle d'accès sur les couches basses interdit toute communication sur les couches hautes avant l'accès au réseau.

Nous avons utilisé la virtualisation pour résoudre ce problème. Sur le lien utilisateur-opérateur nous introduisons une distinction entre les trames de gestion et les trames de données. Cette approche constitue un deuxième canal (virtuel) pour tout accès au réseau. Ensuite, nous excluons le trames de gestion du contrôle d'accès. Ces trames de gestion doivent être traitées différemment. Par exemple, nous proposons de les renvoyer vers une nouvelle entité, le serveur de signalisation (SIGS) mais, conforme à notre séparation de l'interface et de l'implémentation, d'autres réalisations sont possibles. Ce nouveau canal

virtuel peut être utilisé d'une manière restreinte pour permettre la signalisation préalable à l'authentification mais peut également servir pour l'authentification même et pour la découverte des services éventuelle pendant la session. Nous avons étudié les différentes approches à une telle virtualisation et, principalement pour des raisons de disponibilité et simplicité, nous avons opté pour l'adaptation du protocole EAP de l'IETF. Dans cette approche le transport s'effectue sur les couches basses mais les données transportées sont à traiter sur les couches applicatives. Nous avons développé la méthode EAP-SIG qui répond aux exigences fonctionnelles du protocole d'accès commun aux services dans le cadre 4G. Nous considérons explicitement la sécurité des échanges dans le protocole développé puisque la possibilité de communiquer avant l'authentification change le modèle de menaces de l'EAP de base. Nous avons en suite implémenté nos idées (client EAP-SIG, serveur SIGS) dans une plate-forme utilisant une infrastructure sans fil opérationnelle (INFRADIO) déployée à l'ENST Paris. Nous avons ainsi démontré l'applicabilité immédiate de nos idées dans les réseaux existants en utilisant les instances, les équipements et les protocoles de transports existants.

I.3.2. Architecture du réseau de l'opérateur 4G

Dans le cadre 4G, le réseau de l'opérateur comprendra principalement plusieurs réseaux d'accès hétérogènes reliés par un réseau cœur commun, implémentant les services des utilisateurs et les plans contrôle et gestion pour les besoins internes.

Une des exigences essentielles dans ce cadre reste la flexibilité de la solution finale. Dépendant de l'utilisateur connecté, le réseau de l'opérateur doit fournir des ensembles de services avec une certaine qualité. Dans l'idéal, les services accessibles ne dépendent pas du réseau d'accès mais seulement du profil de l'utilisateur. La mise en place de ces limites sur le plan de gestion constitue la flexibilité du système du point de vue de l'opérateur.

Nous traitons ce problèmes sur trois plans. Premièrement, nous nous concentrons sur la flexibilité des infrastructures du type IEEE802.11. Ensuite, nous proposons deux approches à la flexibilisation de l'ensemble de services. Finalement, nous étudions le plan contrôle d'un tel réseau de l'opération et proposons une nouvelle architecture du plan contrôle qui assure son passage à l'échelle automatique.

Virtualisation de l'infrastructure physique

Pour répondre à une des exigences primordiales dans le cadre 4G, notamment servir plusieurs types d'utilisateurs par le même réseau, les infrastructures physiques déployées doivent être capable de fournir plusieurs types de services.

Pour étudier cette problématique, nous avons analysé les exigences posées à un réseau local sans fil selon la norme IEEE802.11 dans un environnement du type campus. Un campus est un environnement relativement ouvert, sans une politique de sécurité claire. De plus, une particularité d'un campus est le nombre de populations rencontrées et leur caractère quasi orthogonal. On y trouve des visiteurs à titre occasionnel, des stagiaires et des étudiants, des cadres, des chercheurs invités, des professeurs, etc. Pour cette raison, un campus semble une plate-forme adaptée pour tester les exigences de sécurité posées à une infrastructure dans le contexte 4G.

Le standard IEEE802.11 ne prévois pas de possibilité de négociation des paramètres de sécurité de la couche Lien (L2) par utilisateur. En outre, même si de telles négociations sont possibles avec IEEE802.1X et prévues dans les extensions de sécurité futures selon IEEE802.11i, elles se limitent à la méthode d'authentification et ne comprennent actuellement pas de négociations de niveaux de sécurité en terme de cryptosuite

(protocole de sécurité, etc.). Notre analyse de besoin a montré que les différentes populations nécessitent une diversification plus importante. Celle-ci varie entre un accès libre (sans contrôle d'accès L2, par ex. pour la distributions des informations, pour les visiteurs) et un accès contrôlé avec du chiffrement fort à clés dynamiques sur le lien (pour les permanents). En plus, la qualité de service doit également correspondre au profil de l'utilisateur. De plus, en pratique la plate-forme des services change selon la catégorie des utilisateurs ce qui nécessite un autre routage.

Il est trivial de répondre à ces besoins en installant plusieurs infrastructures en parallèle. Toutefois, en particulier avec les réseaux sans fil qui ne nécessitent aucun contact physique, la virtualisation suggère une solution plus adaptée. En effet, au lieu de multiplier les infrastructures, il convient de construire une infrastructure qui saura répondre à chaque requête reçue d'une manière appropriée, se comportant ainsi comme plusieurs infrastructures à la fois. Grâce au caractère sans fil, cette virtualisation reste invisible pour l'utilisateur. Elle peut être implémentée en utilisant les points d'accès virtuels, i.e. des points d'accès qui servent plusieurs réseaux sans fil en même temps et les traduisent en marques sur le réseau filaire (par ex. du type VLAN IEEE802.1q)

Nous avons installé un réseau IEEE802.11, opérationnel aujourd'hui à l'ENST site Dareau, qui permet un déploiement d'une nouvelle infrastructure (virtuelle) sans que l'administrateur du réseau ait besoin de quitter la console d'administration. Les infrastructures déployées ainsi permettent principalement de définir les cryptosuites disponibles par infrastructure, la qualité de service à supporter, le routage suivant, etc. Actuellement, notre réseau utilise trois infrastructures virtuelles (visiteurs, permanents et administration du réseau) et plusieurs infrastructures de test utilisées par les groupes de recherche à l'Ecole. Le prototype EAP-SIG décrit ci-dessus est également installé dans un réseau virtuel.

Virtualisation des services

Dans notre modèle du réseau de l'opérateur 4G nous présumons qu'une plate-forme de services commune installée dans le réseau cœur de l'opérateur relie plusieurs réseaux d'accès hétérogènes. Cela implique que le réseau de cœur de l'opérateur doit également pouvoir fournir des différents service et cela d'une manière adaptative car le service utilisé doit pouvoir s'adapter dans le temps aux changements du contexte de l'utilisateur, par ex. au changement du réseau d'accès.

Autrement dit, un utilisateur 4G doit pouvoir retrouver le même environnement indépendamment de l'opérateur et du réseau d'accès utilisé de ce dernier. Cet environnement, en terme de services disponibles, leurs qualités, la sécurité des échanges, etc., doit dépendre seulement du profil de l'utilisateur défini dans son contrat. Principalement, l'adaptation dynamique nécessaire peut se faire sur les deux côtés du lien sans fil, i.e. du côté du terminal ou du côté du réseau. En poursuivant ces deux approches possibles, nous avons proposé deux architectures du réseau de l'opérateur 4G.

La première, appelée RESACO, poursuit l'approche de l'adaptation dynamique du réseau de cœur de l'opérateur au profil de l'utilisateur. Dans cette approche, nous avons poursuivi l'approche d'un réseau programmable capable de se transformer dynamiquement en environnement attendu par l'utilisateur grâce à des entités programmables et adaptatives. Notre participation dans ce projet a abouti dans le développement d'un mécanisme adapté de découverte de services prenant en compte les données du profil. Nous avons étudié les protocoles de découverte de services existants en prenant en compte la disponibilité immédiate du logiciel du côté client et avons opté pour une architecture basée sur le concept UPnP. En adaptant l'architecture UPnP à nos besoins nous avons proposé une architecture s'intégrant avec les briques de sécurité, de

filtrage, de QoS et d'adaptation qui résout le problème de découverte de service, évite les messages diffusés dans le cœur du réseau et peut être utilisée avec les clients existants.

La deuxième approche, MMQOS, suit le principe de l'adaptation dynamique du coté client. L'adaptation sur ce côté du lien a en effet des avantages car elle évite les problèmes directement à leur source. Or, respectant le modèle de confiance, le client ne peut pas être lui-même responsable pour le contrôle de son propre accès. Pour cette raison nous avons choisi un module actif appartenant à l'opérateur d'origine et installé dans le terminal du client pour les taches de contrôle d'accès au réseau. Similairement à l'approche utilisée dans GSM, nous utilisons une carte à puce pour contrôler l'accès au réseau. Toutefois, l'utilisation de la carte à puce dans MMQOS dépasse largement la virtualisation de la méthode d'authentification utilisée dans GSM. Dans MMQOS, la carte à puce construit une interface réseau virtuelle et, dans ce sens, est comparable à un module du réseau privé virtuel (VPN). Contrairement à un module VPN, notre carte à puce se connecte toujours au réseau de l'opérateur visité et utilise les services disponibles dans ce réseau. Cette approche est plus efficace car ces services sont plus « proches » du client. En même temps, la carte est capable de compenser pour le manque de services dans le réseau visité en utilisant les services manquants à partir du réseau d'origine. Ces procédures restent transparentes pour l'utilisateur, car suite à la reconnexion de la carte les services sont toujours disponibles sur la carte qui elle implémente la pile TCP/IP. Nous avons proposé une architecture de l'opérateur 4G/MMQOS qui définit quatre phases d'accès. Dans une première phase la carte ouvre un lien vers le réseau de l'opérateur, en utilisant les méthodes de la couche 2. De cette façon la carte masque l'hétérogénéité des méthodes d'accès du client. Dans une deuxième phase la carte recherche les services disponibles dans le réseau visité par le réseau d'accès utilisé. La carte reconfigure ces remplaçants des services (proxy) embarqués pour utiliser les services disponibles localement et les services manquants dans le réseau d'origine paramétrés correctement. La carte construit une liste des réseaux d'accès disponibles en répétant les phases 1 et 2 pour chaque réseau d'accès disponible. Ensuite, dans une troisième phase, la carte propose le choix au client et demande son autorisation (code PIN, etc.). Dans une quatrième phase, la carte ouvre une interface virtuelle pour donner la connexion au terminal du client.

Décentralisation du plan contrôle

Dans le dernier chapitre de la thèse nous proposons une approche alternative à l'organisation du plan contrôle du réseau d'un opérateur 4G. Nous constatons que toutes les solutions de gestion et de contrôle d'accès existantes se basent exclusivement sur des architectures centralisées. Cela représente des inconvénients majeurs quant à la robustesse, au passage à l'échelle et au coût.

Premièrement, l'entité centrale, introduite par les architectures centralisées (comme par ex. SNMP et AAA) représente un *point de faiblesse* concernant tout le réseau déployé. Notamment, si cette entité centrale tombe en panne, le service rendu par le réseau devient non disponible pour tout utilisateur.

Deuxièmement, comme avec toutes les solutions centralisées, une situation de surcharge peut être provoquée par une activité élevée du réseau, par ex. par un nombre trop élevé d'entités (clients, agents, etc.) déployées dans le réseau et soumises à cette même entité centrale. Cette dernière représente alors naturellement un *goulot d'étranglement* et limite le *passage à l'échelle*.

Finalement, les deux problèmes (robustesse / point de faiblesse, passage à l'échelle / goulot d'étranglement) évoqués jusqu'ici sont résolus aujourd'hui par une fiabilisation de l'entité centrale ou bien par une redondance élevée (par ex. par la hiérarchisation).

Malheureusement ces deux approches provoquent une hausse des coûts souvent inacceptable. La fiabilisation de ce système, la complexité de sa gestion et de sa maintenance, les forces et les compétences humaines nécessaires pour ce faire, etc. représentent des compromis très défavorables au déploiement des certaines technologies (telles que IEEE802.11 par exemple).

Un autre point important est l'observation que les approches centralisées ne permettent pas de suivre la croissance naturelle du réseau. On peut résumer ce fait sous le terme général de mauvais passage à l'échelle de tout système centralisé : il est naturellement soit sur- soit sous-dimensionné à un certain moment.

A cause des problèmes identifiés dans les sections précédentes, une approche radicalement différente semble nécessaire. En effet, toute intégration des solutions habituelles résultera dans une centralisation ou dans l'introduction cachée d'un point de faiblesse. Pour répondre aux exigences que nous définissons (telles que passage à l'échelle, tolérance aux fautes, extensibilité du réseau, compatibilité, gestion du réseau et des utilisateurs et performance adéquate), nous abandonnons les solutions classiques. Nous proposons une nouvelle approche qui se base sur l'intégration de la gestion et du contrôle d'accès d'une part avec une infrastructure suivant le paradigme peer-to-peer (P2P) d'autre part.

La technologie P2P récente a été originalement développée pour le partage de fichiers dans l'Internet (Napster, Kazaa etc). Le récent progrès dans la recherche a abouti à des résultats intéressants, ce qui nous permet de songer à utiliser les protocoles P2P modernes dans le cadre de la gestion distribuée et de l'auto-configuration du matériel déployé.

Nous proposons alors d'utiliser directement les éléments du plan réseau (par ex. les points d'accès au réseau) pour la sauvegarde des données relatives au contrôle d'accès, à la gestion du matériel et à leur configuration. Toutefois, puisque les ressources des éléments fonctionnels peuvent être assez limitées, l'idée générale est de distribuer la charge administrative sur tous ces éléments.

Nous avons proposé une architecture pour les réseaux IEEE802.11 basée sur l'idée de distribution du plan contrôle. Dans notre approche, chaque point d'accès 802.11 garde seulement une partie de la base de gestion commune augmentant alors la capacité de stockage de données de cette base, la robustesse du système et permettant le passage à l'échelle. Notre système organise les points d'accès dans une toile P2P. Optionnellement, d'autres entités (telles que serveurs fichiers, serveurs de sauvegarde, console d'administration, etc.) peuvent également en faire partie. La toile P2P permet la recherche distribuée des données contenues dans ce système reparti. Pour implémenter cette approche nous avons opté pour les technologies DHT en utilisant CAN comme concept du réseau overlay.

I.4. Conclusion et perspectives

Puisque aucune technologie actuelle ne permet de résoudre le problème d'accès au réseau dans notre définition, la vue d'un système 4G en tant que système permettant le choix dynamique et flexible des technologies de transmissions est un concept intéressant. En revanche, la maîtrise d'un tel système hétérogène est un défi technologique. Le vrai problème n'est pas le transport brut de données mais l'organisation de *comment* les données sont à transporter par des infrastructures différentes, appartenantes à des autorités différentes. Le dilemme se pose entre l'homogénéisation et la diversité : alors qu'exiger l'homogénéisation complète n'est pas réaliste, préserver toute la diversité manque d'intégration. C'est pourquoi un compromis raisonnable doit être trouvé entre la

standardisation (i.e. homogénéisation) d'un côté et la flexibilité (i.e. diversité) de l'autre côté.

Pour aboutir à un compromis optimal, le système doit être orienté vers l'utilisateur, contrairement aux approches classiques mettant l'accent sur les services ou sur les opérateurs et résultant par conséquent dans un système limité aux services anticipés ou dans un système avec une architecture fermée et les coûts élevés pour l'utilisateur. L'approche orientée vers l'utilisateur implique que chaque service peut être utilisé par tout le monde, indépendamment de la technologie utilisée, de l'opérateur de l'infrastructure traversée et de l'autorité avec laquelle l'utilisateur a signé son contrat. Le contrat de service n'est donc utilisé que pour établir une confiance préliminaire entre l'utilisateur et l'autorité. Si cette contrainte peut être résolue autrement, les contrats de service classiques ne sont plus nécessaires. Cela est vrai par exemple pour un organisme de cartes de crédit agissant en tant qu'opérateur virtuel.

Dans ce travail, nous appliquons les principes du design orienté vers l'utilisateur dans les problématiques différentes des systèmes de communications mobiles de la prochaine génération. Par cette approche, nous ramenons la complexité du système à un modèle d'accès simpliste définissant un utilisateur, un opérateur du service et l'opérateur d'origine. L'utilisateur accède aux services dans le réseau de l'opérateur de service. L'opérateur d'origine est utilisé comme un point commun de confiance. L'opérateur de services agit en tant que réseau d'accès aux services. La position du service lui-même n'est pas déterminée. Cette vue nous permet de réduire la complexité du système par le biais des interfaces. Nous distinguons et traitons notamment l'interface utilisateur-opérateur et l'interface inter-opérateur. Nous appliquons ensuite la virtualisation et le « cross-layer design » pour pallier l'hétérogénéité des réseaux et les problèmes provenant des systèmes multicouches respectivement.

Pourtant dans un système compliqué et hétérogène composé de plusieurs technologies et de différentes autorités, il existera toujours des problèmes ouverts et le besoin pour des améliorations. En plus, puisque nous n'avons pas étudié la quatrième interface (utilisateur-utilisateur), nos contributions ne s'adressent pas à toute une catégorie des problèmes comme par exemple à la qualité de service de bout en bout.

Pour la plupart de nos contributions une poursuite des études et notamment des expériences pratiques dans un environnement opérationnel est nécessaire. Nous identifions notamment les points suivants comme des points potentiellement intéressants pour poursuivre les investigations :

- Le prototypage d'EAP-SIG peut être poursuivi. Ce prototype peut nous aider à comprendre à quel point les caractéristiques héritées d'EAP représentent une limitation pratique. Principalement, EAP-SIG peut être étendu pour offrir une interface virtuelle vers le plan de contrôle du réseau visité. Cette interface pourrait être ensuite utilisée par des concepts existants de découverte de services, comme par ex. UPnP, etc. Le grand avantage d'une telle approche serait l'intégration simpliste. Toutefois, nous croyons que le canal de signalisation établi par EAP-SIG doit être intentionnellement limité puisque cette limitation rend les attaques par ce canal moins attractives.
- Alternativement, on pourrait considérer le développement d'un nouveau protocole pour le transport générique de la signalisation. Un tel protocole n'aurait pas la compatibilité immédiate comme disponible avec EAP-SIG. En même temps, ce protocole pourrait être basé sur un modèle de communications autre que client-serveur et permettant alors une sollicitation du côté serveur. Les efforts poursuivis depuis récemment dans le groupe de travail IEEE 802.21 semblent intéressants dans ce cadre.

- Notre nouveau modèle d'accès avec les deux canaux indépendants pourrait être utilisé pour une découverte plus souple du réseau d'accès et le réseau de l'opérateur. Par exemple, le canal de signalisation établi par un réseau d'accès pourrait être utilisé pour la découverte de l'autre réseau d'accès et/ou les paramètres des autres réseaux d'accès. Cette approche semble intéressante par ex. pour la coopération UMTS/WLAN.
- Depuis notre travail dans RESACO, de réels points d'accès 802.11 avec un Linux embarqué sont apparus sur le marché. De ce fait, notre approche pourrait être implémentée et testée dans un environnement opérationnel.
- D'une manière analogique, le progrès dans le design des cartes à puces et le développement des nouveaux types incluant le support pour la technologie Java et la pile TCP/IP ouvrent une possibilité d'implémenter un prototype. Ce prototype pourrait confirmer les valeurs de performances obtenues analytiquement.
- Bien que les deux architectures proposées, RESACO et MMQOS, se basent sur l'intégration des mécanismes et protocoles standards supposés avoir des performances adéquates, les architectures pourraient profiter d'une étude de robustesse et de passage à l'échelle dans les scenarii supportant la mobilité de l'utilisateur. Cela pourrait être accompli par une plate-forme de test décrite ci-dessus ou par des simulations.
- COMPASS est un concept intéressant qui vaut une étude supplémentaire. Bien que l'approche soit valide, nous nécessitons des valeurs de performance des réseaux overlay dans le cadre des réseaux locaux. Puisque les implémentations du concept DHT existent maintenant, une vraie plate forme pourrait être développée par ex. dans un environnement 802.11. Une alternative aux technologies DHT devrait être considérée.

Finalement, il semble qu'il ait un fort potentiel non utilisé dans l'exploration des effets synergétiques entre les contributions. Cela comprend par exemple l'usage d'EAP-SIG dans RESACO, MMQOS et COMPASS. Cet usage est principalement directement supporté par le design ouvert de ces architectures. En plus, le développement récent des smartcards EAP ouvre des nouvelles possibilités. Une architecture complète utilisant les approches les plus intéressantes (EAP-SIG avec MMQOS ou RESACO) avec COMPASS dans le plan de contrôle est une proposition très novatrice et serait intéressante à explorer de manière approfondie.

Abstract

The future of the business and personal mobile communications will go far beyond the existing voice (2nd generation) and emerging limited data (3rd generation) communications. We pursue the idea of the next generation global telecommunications system capable of providing a highly adaptive and context-aware access to services all over the world. We discuss different problems with the existing wireless and mobile systems and give an argumentation for a heterogeneous, technology-opportunistic, user-centric 4G vision combining the economic and technological advantages of different transmission technologies.

We show that various new technological challenges are imposed by this specific 4G vision. In particular, we identify the logical service access control and the associated user and infrastructure management as major challenges and critical issues. Indeed, to protect the deployed infrastructures and simultaneously provide a user-friendly anywhere, anytime, always-on service access, new solutions seem indispensable. Both the current infrastructure-centric access control technologies and the operator-oriented, fixed subscription models allow access to the services in the contractually limited service infrastructure of the home operator. In our 4G vision, service access should only be limited by user's profile that alone will represent the contractual situation. In particular, it should be completely operator- and infrastructure-independent and context-adaptive regarding the used mechanisms and scalable, rapid and easy-to-manage from the control point of view.

In the context of the general logical access problem, we present multiple specific problems related to heterogeneous systems. We discuss the existing solutions and demonstrate the shortcomings of both the technology-bound low-layer and the overlaid

high-layer access architectures. Applying new paradigms, we propose a novel low-layer solution which virtualizes the access control mechanisms without pushing the exchanges to the higher layers. We show how this solution can be principally implemented in our 4G vision. We present a functional implementation for a wide-spread wireless local area network technology.

Regarding the user and network management, we first study the possibilities of the centralized access control and discuss issues and possible improvements in the 4G scope. We notably underline the need to decouple the internal organization and management architecture from the user interface implementation. We propose two different approaches to the internal organization of the provider networks covering various network management aspects. We then present our new proposal which applies the peer-to-peer (P2P) paradigm to the control plane. We show the advantages and the open issues with this approach. We then discuss how it can be reasonably implemented in a given infrastructure on the example of a wireless local area network. We also discuss the coexistence of the proposed distributed and classic centralized access control architectures in roaming scenarios.

Finally, we identify some open topics and give an outlook to the future research.

Table of Contents

Table of Illustrations

List of Tables

C H A P T E R I I

Introduction

The proliferation of the Internet technologies and the ongoing world-wide computerization create an urgent need for ubiquitous and convenient mobile communications for both private and business users.

Driven by the IP proliferation and the started convergence process between the existing wireless access technologies, so-called All-IP architectures have been proposed as a unified overlay building the core of the future 4th Generation Mobile Networks.

Because of their obvious appealing advantages like e.g. complete independency of the underlying transmission technologies, All-IP architectures have so far received a tremendous interest in academic research. However, some problems like e.g. delays are inherent to the layered overlay networking and can not be easily resolved.

In this Chapter, we introduce the basic problem areas of the All-IP architectures and difficulties with the alternative approaches. We present our contributions in the scope of 4G networking and describe the organization of the thesis at hand.

II.1. Background and Motivation

The developed world is currently moving towards a fully connected Information Society. The Information Technology sector (IT) is rapidly developing. The Internet has become one of the key communications infrastructures with email service largely outperforming the postal service for business communications and the Web (WWW) evolving to a combined world-wide market place and information source.

In this scope, the wide availability of portable terminals such as personal digital assistants (PDA), digital notebooks and smart phones opens new possibilities for personal and business communications. This is amplified by the increasing dependency on being informed in business affairs and the wish to stay connected in private life.

Wireless communications can provide a truly ubiquitous access to the global data exchange networks like the Internet. With the tremendous success of 2G, wireless technologies have proven to be economically reasonable and technically reliable enough in a large-scale deployment. 2G technologies like GSM have also enabled provision of an almost ubiquitous telephone service in the developing countries.

Generally, users expect support for various services with respectively appropriate quality at lowest price possible. This includes e.g. always-on dynamic rate broadband data access services and session-based constant rate multimedia services. Network providers expect manageable, reliable, highly flexible and cheap infrastructures which can be configured to correctly provide the requested services to different users.

The advent of wireless communications has so far produced a panoply of wireless access networks providing different services like e.g. inter-device communications, classical local networking, Internet access, telephony, etc. Different existing wireless access technologies cover a broad spectrum of telecommunications services in terms of provided data rates, coverage areas, assured quality, mobility support, etc. The most prominent representatives today are cellular technologies like GSM and UMTS, wireless metropolitan and local networking e.g. IEEE 802.16 and IEEE 802.11 and wireless personal networking e.g. Bluetooth, ZigBee, etc. These technologies are being steadily revised and new definitions are being added that improve services or provide new ones. Moreover, the current rapid development of the numeric communications will surely culminate in the development and provision of new access technologies providing a constant evolution of the technical parameters.

Until now however, the development of the wireless access technologies has been mostly driven by the technical requirements of specific services. For instance, 3G cellular technologies can provide a reliable end-to-end multimedia service (telephony, videoconferencing, etc.) but exhibit several weaknesses for data access (restricted connectivity, long link delays, limited data rate, etc.). Simultaneously, the wireless local area network standards provide a convenient means for data access but can not be currently used for services requiring guarantees. Generally speaking, no existing technology can cover the needs of all user classes.

As a result, the next generation of wireless technologies is believed to follow a less technology-oriented and a more opportunistic user-oriented approach. Indeed, the ongoing miniaturization and the constant progress in handset development principally allow integration of different wireless interfaces in a single device. Thus, instead of trying to combine different service classes within a single access technology, the research reorientates towards the provision of services over multiple technologies to the same user.

Therefore, the 4G is expected to be a heterogeneous system combining different access technologies belonging to different authorities (e.g. access or service providers). This 4G view has several appealing properties. First of all, in that definition the system integrates the existing technologies and thus provides for infrastructure and resource reuse. Second, using the most appropriate technology per used service, it can theoretically achieve a more optimal infrastructure usage and user satisfaction. Third, it presumes a multi-provider context giving providers the demanded flexibility in their infrastructure choice and allowing for back- and upward compatibility.

However, the handling of heterogeneous systems bears new problematic aspects. The design of such systems in light of both technological and practical constraints reveals to be much more complicated than the sum of the complexities of used components. This can be studied from the provider or from the user view respectively. Also, for efficiency and interoperation reasons certain criteria have to be fulfilled by the transmission technology itself. The interoperation should not however result in the full homogeneity and should preserve the flexibility for the providers and an attractive service for users. Whether a particular wireless transmission technology is suitable for the usage in this 4G definition depends on the degree to which this given technology fulfills the requirements implied by the problems presented in following.

II.2. Problem Statement

Currently, the only existing technical implementation proposal for such an integrating 4G definition is an overlay-based approach unifying different transmission technologies on higher layers. The overlays are popular because they naturally minimize the adaptations required in the underlying transmission technologies. Indeed, in a overlay solution driven to its extremes the raw data transport is the only requirement of the lower layers. In light of the proliferation of Internet technology, the most popular current overlay solution is All-IP.

All-IP systems are principally capable of data transport in heterogeneous architectures. Yet the problems of the All-IP systems lie elsewhere, namely in the mechanisms around the data transport, i.e. network access, connection establishment, service access, etc. For reasons which we detail later, it reveals to be advantageous (e.g. more efficient) and sometimes indispensable to take into consideration the native mechanisms proposed by the transmission technologies. This is however problematic because this is exactly where the heterogeneity lies.

The handling of the changing and dynamic network environment spanning over several potentially unstable links represents a challenge from both system and handset design perspectives. Herein, the most typical problem areas are:

- heterogeneous security, including multi-entity trust relationships, network access security with the implied reliable network and service discovery, service user part auto-configuration and secure service access with roaming,
- management of the heterogeneous access networks,
- quality of service (QoS) enforcement over heterogeneous networks.

A consequent system-wide definition should allow the establishment of appropriate security, QoS and management levels independently of the currently used access networks. The problem is that the security, management and QoS services provided by the access networks can differ in their very principles.

II.3. Accomplishments and Contributions

In this work, we study the different aspects of 4G architectures assuming a general All-IP idea i.e. a common IP-based data transport over a heterogeneous system. We present our system vision, define our system access model with its main entities and authorities. We then analyze the identified problem areas assigning them to one of the three interfaces in the system:

- user-provider interface has to provide a coherent signaling over heterogeneous links, including security and environment related parts;
- provider internal interfaces need to support the management of the installed entities;
- provider-provider interface has to support mutual roaming agreement verifications.

The applied interface-based system analysis permits the application of the virtualization techniques known e.g. from higher programming languages in the system design. The virtualization decouples the interface design from the internal implementation and can help to overcome the encountered heterogeneity of the access technologies.

We study 4G integration with a practical accent on the popular IEEE 802.11 technology. Our 4G vision is loosely based on All-IP. Analyzing existing 4G proposals, practical constraints with the heterogeneity, handset design and anticipated user service requirements like mobility, QoS, etc. we come to the conclusion that a limited usage of the technology-dependent mechanisms can be advantageous in different 4G areas. We notably justify the requirement for the usage of the technology-dependent security mechanisms.

Studying the problem area on different identified interfaces and using new approaches like virtualization, dynamic adaptation, P2P networking, etc. we make the following contributions:

- We analyze the trust relationships in our system architecture and typical roaming possibilities in IP-oriented architectures. We optimize the provider-provider exchanges in the case of a user roaming access minimizing the number of inter-authority messages and the associated roaming delay.
- We study the adaptation capabilities of the access networks at the example of a 802.11 network. Using a fully operational 802.11 network designed according to our concepts, we show how a highly adaptable 802.11 network can be deployed in a campus environment using virtual access points.
- We analyze heterogeneity implications in the network access phase. The identified issues include network discovery, access security and service discovery. We propose an alternative access model implying two distinct logical channels. We propose a virtual homogenized signaling on the second layer. This approach can provide more than a purely IP-based, technology-agnostic convergence but establishes noticeably less requirements than a full technology convergence i.e. a full homogenization.
- We show a possible implementation of our virtual signaling concept in modern IEEE 802.11 wireless local area networks (WLAN) using modern access control measures. We develop Extensible Authentication Protocol Signaling (EAP-SIG), an easy-to-implement generic signaling protocol generalizing WLAN's access control architecture. Our signaling does not need any changes in the access points and can thus be used in the deployed networks. We prove it by installing our prototype implementation in the operational 802.11 network on our campus.

- We apply our virtual signaling concept carrying out a case study of a classical layered IP micromobility integration within a modern 802.11 WLAN as opposed to micromobility integration using EAP-SIG. We show how our approach can help to minimize handover delays and avoids unnecessary double signaling between different implied entities.
- We study the internal organization of the network, control and business planes of the provider networks within our global system vision. We develop a flexible provider network architecture capable of dynamically adapting to the profile of the requesting user. We develop prototype equipment based on IEEE 802.11 technology interacting with the control plane and proposing personalized services to each user.
- Using terminal-inserted smartcards as a provider-owned control element, we develop an alternative approach to the organization of the internal provider architecture. Our new architecture features environment stability for users and extensive control possibilities and optimized service access for providers.
- Finally, we address the organization of the control plane of the internal provider network. We study the shortcomings with the current mostly client/server and agent/manager based, centralized approaches. We develop Configuration Management P2P Access Security System (COMPASS), an original control plane architecture leveraging the recent advances in the peer-to-peer networking. COMPASS provides both a decentralized access control and an integrated configuration and device management of the deployed 802.11 access points.

II.4. Organization of the Thesis

This thesis focuses on the logical network access problem in heterogeneous 4G environments and explicitly treats the associated provider network architectures and user and network management issues. This document is organized as follows:

- In Chapter III, we give a general introduction to the wireless technologies justifying their usefulness for the solution of the access network problem in the scope of world-wide IT services provision. We synthesize the recent impressive technological progress in the development of the wireless technologies and underline different problems inherent to or implied by the wireless technologies.
- In Chapter IV we concentrate on future mobile networking. We analyze the existing wireless technologies and conclude that no currently existing technology is capable of covering the whole spectrum of service needs and scenarios. We thus give an argumentation for the next generation of mobile networking and establish a roadmap. We then present the recent research in the areas of 4G, 4G terminals, 4G mobility and 4G security identifying several still open subjects.
- Chapter V is dedicated to the 4G network access. We present our user-oriented 4G vision identifying the main entities and actors. We then define the system access model. We discuss different issues with the logical network access. This includes user roaming and the associated optimizations on the provider-provider interface. We also identify an insufficient interworking on the user-provider interface and introduce EAP-SIG, our approach to a generalized signaling.
- In Chapter VI, we address the organization of the internal provider architectures. We propose two alternative adaptive architectures providing a stable user service environment and simultaneously preserving the flexibility for the provider to decide about the internal realization and organization of services. We also study the organization of the control plane proposing a novel alternative architecture.

- Finally, in Chapter VII we give an outlook to the future work and identify several still open issues.

The complete bibliography and the terminology index can be found at the end of the document. Implementation details on the implemented concepts and prototypes can be found in the appendices.

C H A P T E R I I I

Wireless Telecommunications

In this Chapter, we first present the access network problem and show that it is the main barrier delaying the proliferation of ubiquitous data access and constituting the current digital divide.

The access network problem motivates the further development of digital wireless technologies. We proceed with an overview of modern telecommunications from the perspective of the digital wireless transmission technologies. We classify the existing and the emerging wireless technologies and introduce several current hot topics with their respective terms and definitions.

We notably give an overview of various problems inherent to wireless links. We then present several issues related to wireless services, including mobility and security. We show that the resolution of these issues represents a technological challenge increasing the engineering complexity of the wireless technologies.

III.1. Access Network Problem

III.1.1. Wireless Technologies against the Digital Divide

In 2000, the Secretary-General of the UN, Kofi Annan, repeated the original 1994 remark of Greg LeVert of MCI saying in [1]:

> *"Yet half the world's population has never made, or received, a telephone call. This week, our thoughts should turn to practical ways of connecting the people who need it most."*

With the deployment of the second generation of cellular phone systems (2G) starting in 1992, affordable portable telephones became a reality. In the following years, 2G equipment was developed and deployed all over the world. In 2002, two years after Annan's remark and ten years after the introduction of the first 2G systems, the phrase seems to have been proven wrong [2]. The 2G systems appear to be one of the "practical ways" so much sought after. According to the statistics of the Telecommunication Development Bureau of the International Telecommunication Union (ITU) [3][4], between 1995 and 2001 the number of cellular subscribers grew from 91 to 946 millions (156% annual growth rate, as compared to an average annual growth of 7.3% for wired lines). Consequently, the number of wireless subscribers has exceeded the number of the installed wired subscriber lines in most western European countries by 2002. The cellular growth rate in the developing countries has typically been much higher than that of the developed countries: according to [2], by the end of 2000, there were 25 developing countries where cell phone users made up between two-thirds and nine-tenths of all connections.

Given these amazing growth rates, the 2G systems are an efficient technological way to mitigate the so-called *digital divide problem*, i.e. the technological gap between the developed and the developing worlds. At least from the point of view of voice telecommunications, 2G systems are being extraordinary successful. 2G systems have finally brought telephony to the third world countries. Today, even in the poorest cities of the world one finds 2G coverage, typically provided by the European Global System for Mobile Communications (GSM) technology: GSM accounts for over 1.1 billions subscribers across more than 200 countries of the world [5].

Two years after the first deployment of the commercial 2G voice networks, the commercialization of the Internet started. This commercialization has resulted in a rapid development of a world-global information and services network in the following years. While one can discuss the sense and the non-sense of some existing offerings, nobody would honestly contest the general usefulness of the Internet for education and communication purposes. This claim is proven by the enormous success of the World-Wide-Web (WWW) and the provided electronic mail (email) services. In the WWW, the instant information availability and the amazingly efficient search engines can provide a faster and easier information access than classic libraries. For personal and professional communications, email outperforms the usual mail services in terms of flexibility and speed. Since ten years both services are widely accepted in academia. Since at least 5 years these are virtually ubiquitous in business communications.

The Internet access enables active participations in world-wide discussions (divers news groups, discussion forums, etc.), instant information access (WWW), personal communications (email, Instant Messaging) and, finally, the connection to the established communication means of the developed world (e.g. over Voice-over-IP, SMS and pager

portals, etc). Through these and other services the Internet provides an interesting alternative to such classic services as broadcast mass media, libraries and post. These latter presume a highly developed local infrastructure and richly equipped facilities – assumptions that simply do not hold in most developing countries.

According to the CIA's World's Fact Book 2004 [6] and the ITU Internet Usage statistics for 2003 [7], only about 10% of the World's population are Internet users. Figure III-1 illustrates the current Internet usage per capita per country based on the ITU data on 56 nations[1] [6][7]. As can be seen, the Internet usage ratio decreases from the right (40%-60% Internet users) to the left (less than 10%). Using the Internet availability indicator, this figure empirically demonstrates one of the newer aspects of the digital divide problem.

Hence, the usage of the new Internet data services seems to constitute a new barrier. Without taking into account the social, educational and financial factors, i.e. from the technological point of view, the availability of an Internet connection depends on three main factors. First of all, it requires suitable terminals. Second, it requires national or commercial broadband backbone connections to the international Internet. Finally, the last barrier is a locally available access network connecting the terminals to the backbone and capable of an efficient user data transport from and to the Internet.

From the point of view of technology, Internet-capable terminals exist en masse. Almost any personal computer, laptop, notebook or a similar device less than 5-8 years old can be used today as an Internet access terminal. Additionally, the development of more powerful portable devices like so-called sub-notebooks, digital assistants, smart phones, etc. leads to the full convergence of mobile and personal computing. Although the terminal availability remains a serious financial problem in the developing countries, the price battle on the hardware markets and the rapid development of the information technology (IT) sector cater for a high renewal rate in the developed world. This provides a rich source of second-hand terminals.

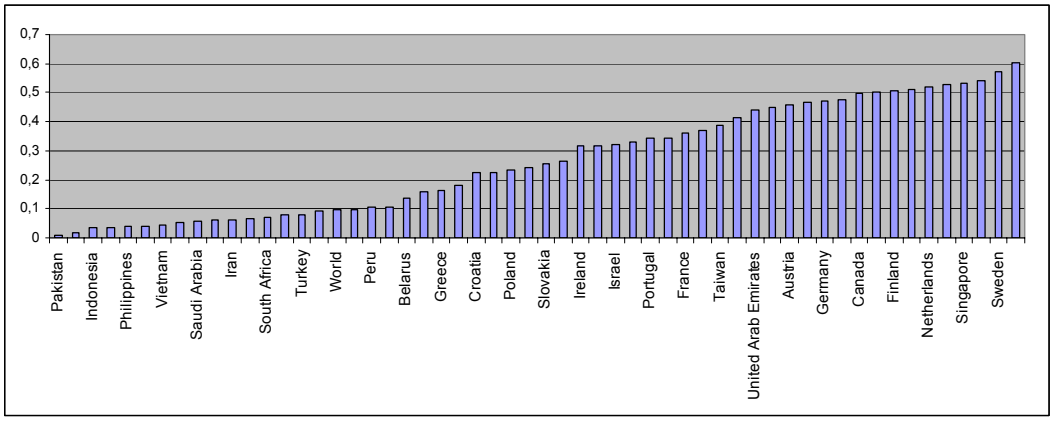


Figure III-1 Number of Internet users per capita

[1] To avoid data instability, only countries with more than 1 million Internet users (10^6) have been considered. The illustrated data cover 77.5% of the World population.

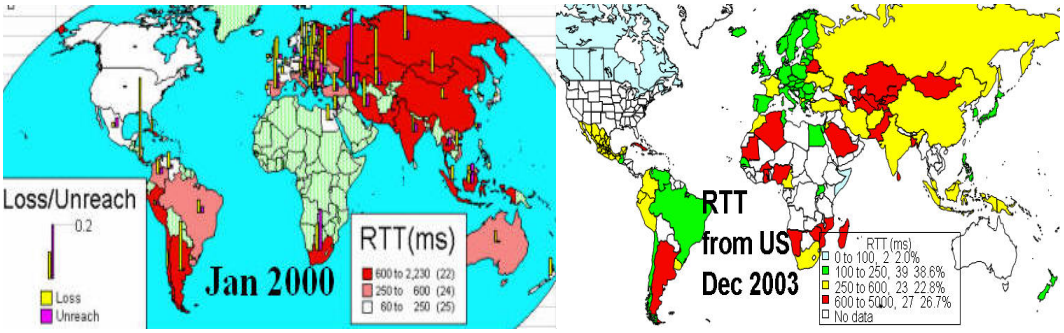


Figure III-2 Development of the backbone links from 2000 to 2004

Regarding the available Internet backbone, the high-latency geosynchronous satellite connections are still very important for countries with a poor telecommunications infrastructure. However, according to the recent studies of the digital divide [8], these links are being steadily replaced by landline connections (typically by fiber), primarily to provide better latency properties i.e. better round trip time (*RTT*) values. The quality of these connections is improving steadily each year: 40-50% for packet loss, and 10-20% for RTT values. Figure III-2 [8] shows the development of the backbone links. It can be seen that the number of red-marked countries (very poor RTT of 600ms and more) has much decreased. Thus, the necessary technology is available and seems to be steadily increasing in the developing countries.

In spite of the amazing backbone development, the big difference in terms of Internet availability between the developing and the developed countries seems evident. The terminal availability does not represent any technological barrier. Thus, a suitable and cheap data access network technology seems to be one of the key technological barriers constituting the digital divide.

III.1.2. Wireless Technologies for a Ubiquitous Data Access

In the developed world, the broad Internet acceptance and the business demand are the major drivers for the creation of the new *anytime, anywhere, always-on* paradigm. It calls for gapless coverage, fast and global user mobility support and intelligent pricing mechanisms. However, it also raises economic, technological and security considerations.

First, the operators and the users are interested in cost reduction of the data access provision. Current per-minute or per-byte pricing schemes are hardly suitable for the always-on access: data transmissions do not necessarily have any call definitions (which can be accepted or rejected). It is thus generally impossible to control the personal costs of staying connected. The access provision needs to be easier for users and more flexible for operators.

Second, the technological properties (coverage, mobility, cost) and the ease of use imply a wireless transmission technology as an enabling factor for the always-on data access provisioning. Indeed, the realization of the anytime, anywhere, always on is hardly imaginable at all with wired technologies. The used wireless access network technology has to reliably provide very different services and, in particular, high data rate and low-latency data connections. It is thus expected to be highly versatile, powerful and cheap at the same time – obviously representing an almost utopian requirement ensemble.

Moreover, the requested reliability and the associated economic and personal risks need to be explicitly addressed. This calls for mature and manageable security.

The observed slow process of always-on data access provision in the developed countries can not be explained by missing terminals, weak infrastructures or lack of interest. In contrary, it seems that the terminal market is flooded by Internet capable terminals of all genres. Highly developed wired and wireless infrastructures are widely available. The interest is steadily growing with the Internet usage, recently being additionally motivated by the ongoing consolidation of the online businesses and new very popular online offerings in the multimedia sector.

We think that the core problem in the developing countries is the lack of a suitable access network technology. Different technologies exists and some are already in use [9][10]. However, these are either too expensive (too high deployment cost for operators, inappropriate access pricing schemes for users) or their usage is quite limited because they are technologically inadequate (not manageable, insufficient service, too cumbersome) or even both.

III.1.3. Access Network Problem

As was discussed above, a mobile data access technology seems to be the missing part both in the developed and the developing worlds. There currently seems to be no suitable technology to economically and efficiently provide data access in every country.

The design, development and provision of a sufficient access network is a technological challenge which has to cover the access network but also different core network aspects. One particular aspect of the access network problem – fixed broadband data access – is part of the general *last mile problem*. Current solutions are based on the classic telecommunication infrastructures (e.g. telephone or cable TV networks), reusing these for data transport. Technologies like xDSL provide broadband data connections over standard telephone lines. ADSL achieves good downlink data rates of about 2-15Mbps per user with low link latencies (typically under 30ms). These technologies are likely to be further developed in the next future. Nevertheless, such solutions are not suitable for the necessary anytime, anywhere, always-on data access in the developed world. Neither are they applicable in the developing countries because of the inherent lack of reliable infrastructures. Based on the positive world-wide deployment experiences with wireless networks (e.g. with 2G for voice telecommunications), much is expected in this matter from the wireless technologies.

We thus want to give an overview of the available and emerging wireless technologies, discussing their properties and the appropriate pros and cons. We then show why the existing wireless technologies can not resolve the access network problem. We then discuss the new paradigms in wireless telecommunications research and development.

III.2. Digital Wireless Technologies

III.2.1. Classification of Wireless Technologies

Originating from their very different applications, the demand for wireless transmission support has so far produced a large palette of wireless technologies reaching from the global satellite-based communications to the tiny sensor interconnections.

Available wireless transmission technologies can be assigned in different categories according to such criteria as provided data rate, coverage, medium access, used frequency bands, etc. Here, we distinguish the wireless technologies according to their purpose.

Figure III-3 classifies several already available or still in development wireless technologies in respect to the provided data rates (in kbps) and the mobility level. We distinguish five mobility levels (with "indoor" and "outdoor" describing the scope and not the environmental properties):

- Indoor-fixed
  This describes very limited-range technologies where the mobility concerns are negligible because of the size of the coverage area. Typical representatives of this category are wireless personal area networks (WPAN) such as Bluetooth [11] with a typical coverage area of not more than 10m. Other technologies in this category are ETSI HIPERPAN , IEEE 802.15 [12] or ZigBee (802.15.4) [13].
- Indoor-walk
  This comprises limited-range technologies supporting slow mobility within a limited area. Typical representatives are wireless local area networks (WLAN) which provide a local (intra-domain) slow mobility support and a coverage area of approx. up to 100m diameter per access point. The most popular WLAN technology is IEEE 802.11 [14] with its different transmission technologies 802.11b, 802.11g and 802.11a. Other technologies include HomeRF and ETSI HIPERLAN/2 [15].

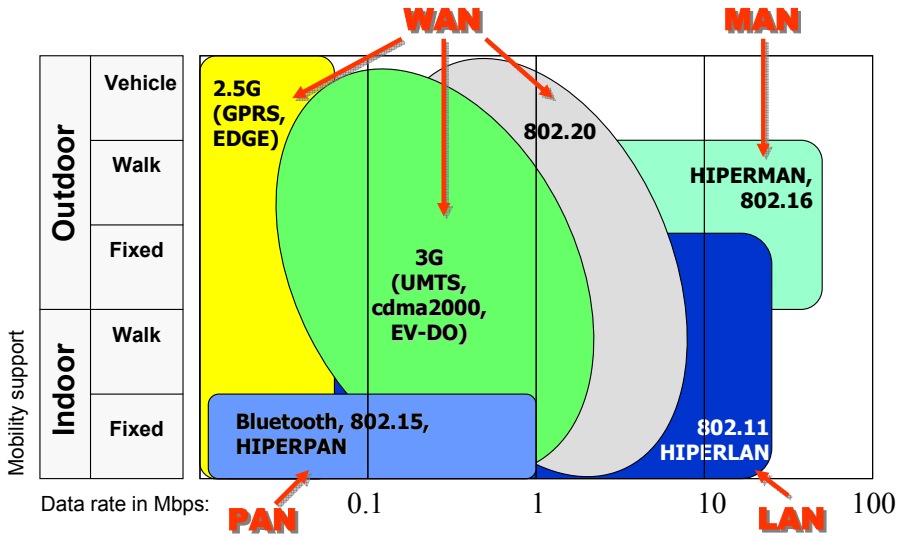


Figure III-3 Classification of wireless networks by data rate vs. mobility support

- Outdoor-fixed
  This comprises wider range technologies which only provide mobility within the coverage area (typically under 500m free movement). Typical representatives are wireless local loop (WLL) technologies such as fixed wireless metropolitan area networks (WMAN), e.g. IEEE 802.16 [16] or ETSI HIPERACCESS (HIPERMAN) [17]. This also comprises outdoor WLAN installations since their coverage area can be extended by increasing the number of the installed access points.
- Outdoor-walk
  This comprises wider range wireless technologies providing true outdoor-scope slow mobility support (more than 500m, with low speed limits). This is a typical usage scenario for a mobility-supporting WMAN providing connectivity within the whole city area. IEEE 802.16e [18] addresses mobility issues in a metropolitan scope.

- Outdoor-vehicle
  This category includes wide range wireless technologies (potential range of up to 30km) providing a true, fast-velocity mobility (up to 250km/h). It includes wireless wide area networks (WWAN) such as the existing cellular technologies (GSM, GPRS, EDGE [19]; IS95, 1X [20]), emerging 3G (UMTS [19], cdma2000 [20]) technologies and the new, planned IEEE 802.20 standard [21] for mobile broadband wireless access (MBWA).

Note that the given data rate is only one possible technological pointer. Alone it is usually not sufficient to make decisions on the suitability of a technology for a planned service. Other characteristics such as provided quality of service classes, assured latencies and latency variations (*jitter*) should also be taken into account. These and other technological factors (e.g. co-existence with other technologies, deployment costs and requirements, spectrum license, etc.) constitute the main reason why different proposals exist for every discussed category producing overlapping areas in Figure III-3.

III.2.2. Wireless Technology Development

Started in 1970 with simple medium access techniques such as ALOHA [22] used in the first digital packet radio services etc., the digital wireless technology was developed over years to support different coordinated multiple access techniques such as FDMA, TDMA, CDMA or distributed multiple access techniques e.g. DAMA, MACA and CSMA [22].

Today, the wireless sector is one of the most rapidly developing IT sectors. Figure III-4 illustrates the approximate observed development speeds of different IT areas contrasting microprocessor, memory size, network switching, local access and wireless technologies [23]. These factors mainly depend on the transistor density, microprocessor speed (brown line) and memory size (blue line) respectively. The latter are known, within some constraints, to follow the Moore's Law [24]. Whereas in the logarithmic scaling of Figure III-4 these thus grow linearly, the short range radios (red line) have been outpacing this development speed for more than 8 years. This amazing development speed suggests the integration of short range radios in the next generation wireless systems.

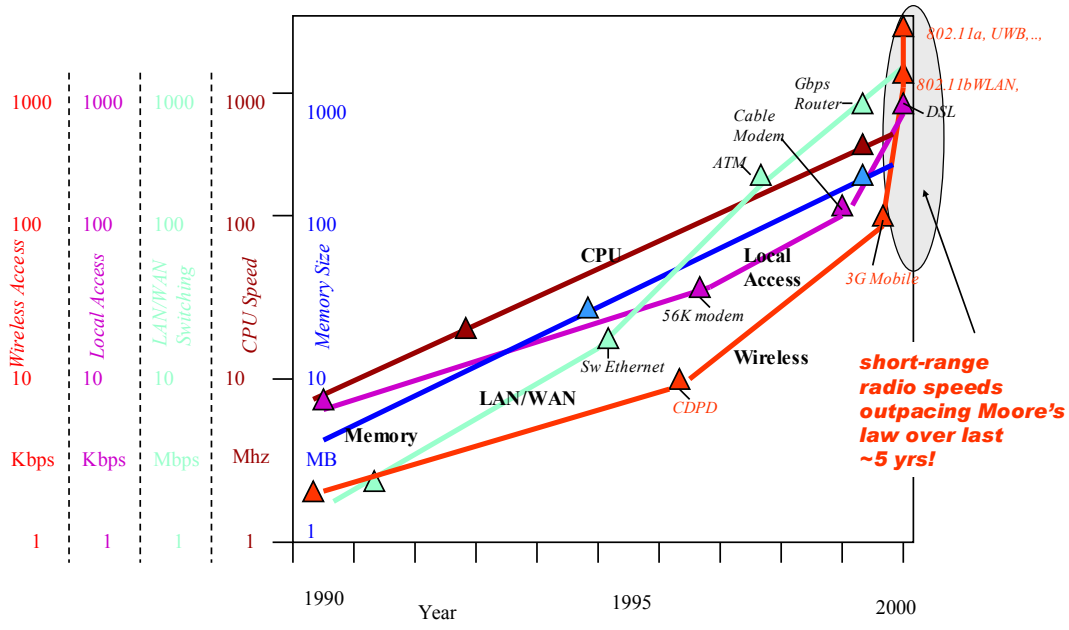


Figure III-4 Development of short-range radios vs. Moore's Law

In the ISO/OSI reference model [25], the wireless transmission technologies (i.e. independently of management and supporting core network functions) occupy the lower two layers (Figure III-5), exactly as their wired pendants. However, to fulfill the wish for higher data rates and more transmission stability in spite of several characteristic wireless link issues, a strictly layered abstraction proves too inefficient. As a result, the wireless network design often requires an introduction of different intermediating or spanning layers. In a general case, a consistent cross-layer design is appropriate from the beginning on.

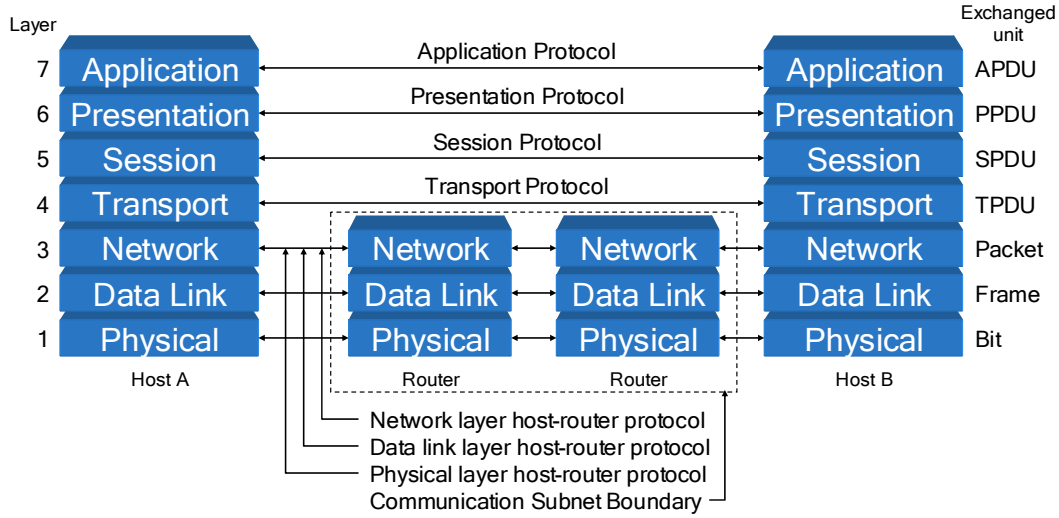


Figure III-5 ISO/OSI reference model

The cross-layer design [26] permits new functional steps in the design process and thus enables new solutions (e.g. dynamic choice of a medium access technique in the Data Link layer depending on the Physical layer information, definition of a Transport layer optimized for coping with higher bit error rates of wireless links, consistent incorporation of security functions in different layers, etc). However, the cross-layer design introduces an additional heterogeneity through new explicit dependencies (dedicated layers, proprietary interfaces, etc). It also demands higher engineering efforts than a classic layered approach since the layered design does not hide complexity by subdividing the problem pile.

We discuss some of the inherent wireless link issues in the next section.

III.3. Wireless Link Issues

Because of the physical medium properties, the wireless links naturally suffer a bigger transmission error rate than the wireline transmissions. They thus require complex error correction and redundancy schemes and typically result in lower performance compared to their wired counterparts. It is however only a part of the problem since efficient algorithms exist and are well-known. The actual difficulty is the high dynamics of an established wireless link in time. Multi-path distribution, fading and interferences depend on the environment, moving objects, etc. Being unpredictable without complete environmental knowledge, these phenomena make it impossible to obtain a physically stable wireless link. The state of the link rather changes continually, thus providing

statistically distributed quality properties. This raises the need for new techniques such as fast and precise channel quality estimation and an efficient adaptive coding.

Currently evolving at the same speed as the wired technologies, the wireless links are typically about two magnitudes behind their wired counterparts. The following issues are inherent to wireless technologies:

- regulation, spectrum licensing,
- broadcast medium and interference,
- propagation ambiguity,
- multi-path, fading, noise,
- exposed/hidden stations.

These issues are the reason why wireless technologies per se often represent a technological challenge. We describe these issues in more detail.

III.3.1. Shared Medium and Spectrum Licensing

Always using the same medium, wireless technologies require agreements to enable multiple simultaneous technology usages. The spatial scope of such agreements normally represents the scope of the expected usage range, which primarily depends on the used frequency and the emission power. In the beginning of the XXth century, the international (ITU) and national (US: FCC, France: ART, etc.) regulation authorities have started issuing so-called *spectrum licenses*, i.e. frequency range reservations for different wireless technologies.

A spectrum license can be obtained from one of these local authorities. It gives an exclusive (local) right to use the assigned frequency bands. To prevent abuse, the usage of a licensed spectrum band without a license is prohibited by the respective national laws.

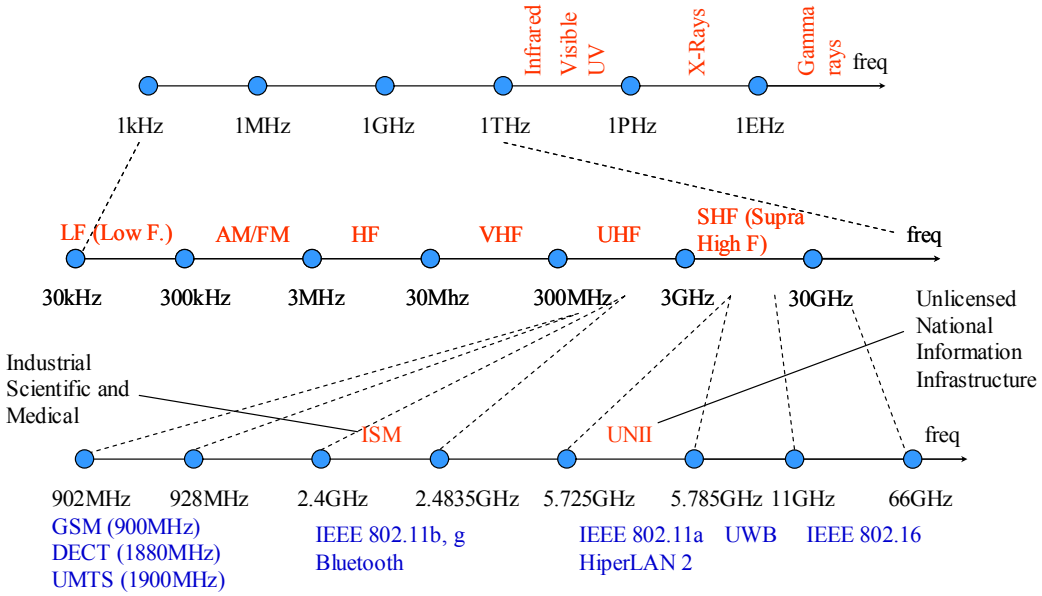


Figure III-6 Current radio spectrum usage

Started with global scope telecommunications technologies like radio, it continues later with television, satellite communications, the cellular communication systems, etc. This

historic process results in a quite complicated frequency spectrum fragmentation in different bands. A simplified and incomplete international frequency range allocation is shown in Figure III-6. The current complete international spectrum reservation is maintained by the ITU-T in the Master International Frequency Register (MIFR) table and can be found in [27].

However, not all bands require a license. To enable an uncomplicated home (e.g. wireless telephony), scientific (e.g. wireless sensors) or general private usage, some frequency bands like e.g. the ISM band or the UNII bands are considered license free. To avoid the mutual impacts of the installed equipment, the emission power of the latter is limited by national regulations. Advanced interference avoidance techniques like e.g. automatic channel selection, dynamic power adaptation, etc. are encouraged by some countries.

III.3.2. Broadcast Nature and Interference

Spectrum licenses can be very expensive [28]. Unlicensed bands are limited and have to be shared with other users. Hence, the wireless bandwidth is always a naturally scarce and precious resource.

Yet, unless directional antennae are used, the wireless transmissions exhibit a natural broadcast character. The so-called *space division multiplexing*, i.e. spatially-parallel usage of the same technology (e.g. on a parallel cable, directional antenna/sectors, etc.) is not always possible in the wireless world.

Fortunately, with most of used frequencies[1], the spatial signal propagation is very limited for all reasonable emission power levels. However, the mutual influence of the installations, the so-called *interference*, goes far beyond the clear signal propagation. Additionally, it has an impact on the neighboring frequency ranges. Unused frequency ranges (gaps) are thus necessary between the used channels, bands, etc. further limiting the available bandwidth. Moreover, the spatial interference is complicated by unpredictable wave propagation.

III.3.3. Propagation Ambiguity

The propagation of electromagnetic waves in space depends on the wave length (frequency), the emission power and the environment. Because of the dynamicity of the latter, the link quality is unstable in time and in place (considering the movements at the site and the changing wave transparency properties of air, walls, used materials, etc.). The propagation of electromagnetic waves is thus indeterministic.

Because of this propagation ambiguity, the interferences can occur at unexpectedly high distances from the sender (where no clean signal can be received anymore). Provoking packet corruption, the interferences are typically perceived as a throughput decrease of the transmission technology because of the necessary retransmissions. For instance, a single microwave oven can disturb the transmission quality of the whole WLAN cell (60m diameter), resulting in the perceived data rate decrease of 60% to 85% [29].

Figure III-7 shows one of the results of our signal measurements of a 802.11 WLAN. This was carried out during the wireless network deployment phase of the French national telecommunications project INFRADIO [30]. In this figure, a single wireless access point equipped with a diversity antenna is located in the 5th floor of the Dareau facility of the

[1] Short waves (3-25MHz), corresponding roughly to the HF band (see Figure III-6), can be refracted from the ionosphere back to Earth, thus being propagated around the Earth.

ENST Paris, at the position marked with a red cross on the building plan (at the left). The signal strength values (in dBm) are measured at different coordinates of the two-dimensional map (by continuous automatic data collection module moved within the building connected to a GPS module). The interpolated collected discrete data result in the 3D-presentation (at the right).

Figure III-7 shows a signal peak at the location of the measured WLAN access point and the fast attenuation following the x-axis, with the signal dropping under -99 dBm (connection loss) within the area marked with an oval (about 11m from the access point).

In vacuum, the signal strength should decrease monotonically, proportionally to the square of the distance from the access point (*inverse square law*). However, as can be seen, the signal strength profile possesses multiple local extrema. In the projection of the equivalent signal values (at the left), the resulting shapes are not circular. This is because of location properties (e.g. obstacles, walls, materials, etc). The shown profile is thus characteristic for this environment.

Still, Figure III-7 does not contain the temporal aspect of the dynamism. Based on multiple snapshot measurements at the same location, it rather represents an averaged signal strength in time.

Additionally to the performance concerns, the propagation ambiguity also raises security concerns as will be discussed later.

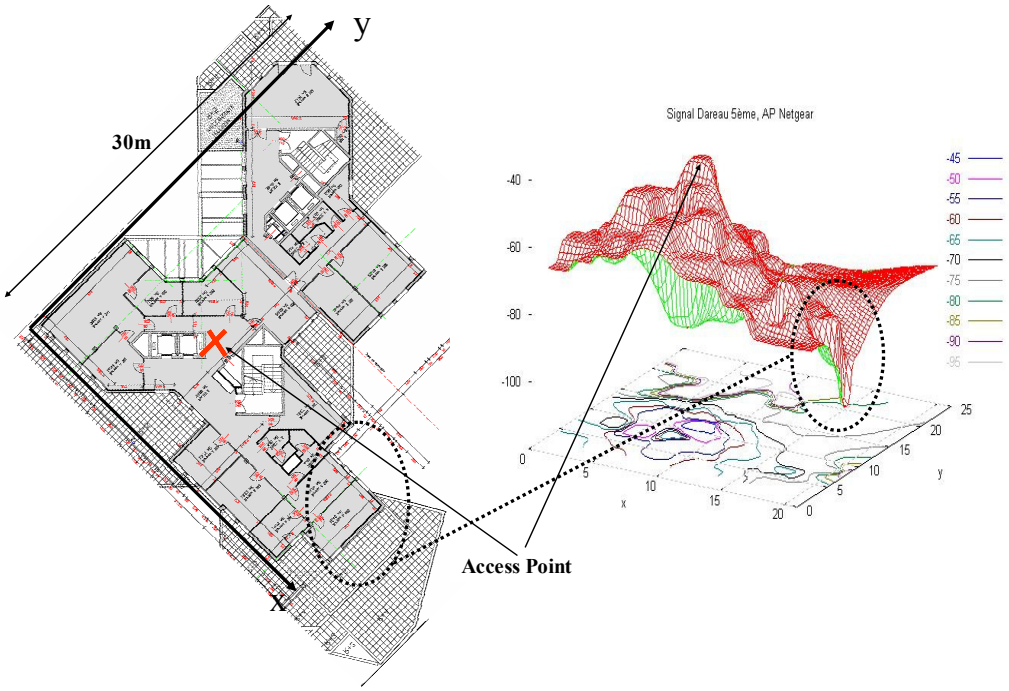


Figure III-7 Measured signal quality of a single 802.11 access point

III.3.4. Multi-path Propagation and Fading

The electromagnetic waves can be reflected from and scattered on certain objects and materials. This results in multiple paths taken by the same signal from the sender to the receiver and creating multiple copies of the same wave (*multi-path propagation*). Since the paths are not necessarily equidistant, the copies arrive at the receiver with a certain time-shift. This leads to an attenuation of a fraction of the signal wavelength (e.g. if the

time shift between two copies roughly corresponds to the half wavelength, the two copies add to zero). This phenomenon is known as *Rayleigh fading*. Such small-scale fading is highly dependent on the environment and on the respective positions of the sender and receiver. It generally accounts for significant power loss.

III.3.5. Exposed and Hidden Stations

In shared medium configurations a collaborative medium sharing method needs to be specified. This method could be coordinated (like SDMA, TDMA, FDMA, CDMA, etc.) or distributed (ALOHA, DAMA, MACA, CSMA, etc.) if no authority is available [22].

The distributed multiple access methods need to address the *collision* problem, i.e. the case where two stations try to use the same medium at the same time. This results in corruption of both signals. Collisions thus decrease the overall performance. To avoid sending on the busy medium, stations typically use a *medium sense* method before accessing the wireless medium. However, it is still possible that two stations start an emission simultaneously. In the wired distributed architectures like e.g. Ethernet, *collision detection* is used to resolve these cases.

In a wireless environment, the medium sense is often problematic because of the fast and ambiguous attenuation of the power level with the distance. Hence, the classic collision detection can not be used. It is possible that the medium is used at the receiver but not at the sender thus not avoiding the collision at the receiver (*hidden station problem*). On the other hand, the medium could be perceived as busy at the sender but not at the receiver thus unnecessarily avoiding a possible transmission (*exposed station problem*).

Such problems need to be addressed by an additional control packet exchange (request to send, clear to send, acknowledgment, etc.) between the sender and the receiver. However, such control packets additionally delay data transmissions and decrease the wireless link performance compared to the wired media.

III.3.6. Synthesis

All discussed wireless link issues have to be explicitly addressed by the wireless technology standard. Hence, the design of a wireless technology represents a technological challenge per se. The discussed “physical” issues are usually covered within the physical layer and the medium access layer parts of the related standard. However, the wireless technology can have an additional impact on other subjects of networking and other services. We discuss these in the following sections.

III.4. Mobility

Mobility is a much discussed topic in the wireless world. Yet, wireless technology does not imply mobility. For instance, when using directional antennae, a wireless link can be thought of as a fixed connection prohibiting any considerable position changes of the involved parties. Moreover, neither does mobility presume any wireless technology. For instance, Mobile IP [31] or SIP [32] mobility technologies can be used with both wired and wireless terminals.

However, there is a considerable overlap between mobility and wireless in practice. On one hand, a lot of wireless technologies provide a natural mobility support within the coverage area. This area typically largely exceeds the radius of a cable connection. For instance, with a cordless telephone, a user can freely move within her home while talking.

On the other hand, mobility definitions presume some dynamism (e.g. terminal mobility presumes a loose connection of the terminal). Even though this dynamism can be achieved with wired technologies, it is primarily with wireless that we experience the convenience of full scale mobility without the limiting cables. For example, WLAN users tend to change the points of attachment much more often than wired users [33].

Thus, using wireless links terminal mobility becomes a much more interesting feature. We thus have to think about combinations of wireless technology and mobility. Mobility support is however a technological challenge. Given that wireless technology is a challenge per se, we have to cope with a complexity increase.

We distinguish different types of mobility according to the following aspects:

- locality of its impact (from where to where, global/local, etc.),
- subject of movement (what, user/session, terminal, network, etc.).

Local mobility, i.e. changes of the points of attachment within some well-defined limits (e.g. LAN/domain/authority), is typically called micromobility, while the term macromobility is generally used for more global, inter-network or inter-domain mobility (e.g. over Internet).

Note that there is no exact general definition of these terms, since their usage depends on the context. For example, the movements within a cellular operator's network are handled by the same authority and the same network. Technologically, this mobility could be considered as micromobility. At the same time, the mobility scope of the terminal covers national scale mobility. Geographically, it thus suggests the macromobility term.

The question to what is mobile also plays a crucial role in the mobility question. *Terminal mobility* presumes the change of the point of attachment of the terminal to the network. It typically occurs when the terminal is moved (by the user) from one network access point to another. It is however imaginable that the access points move relatively to a fix terminal. That is called *network mobility*. It defines the common movement of a group of terminals and the network equipment (e.g. a local network installed in the train passing by). Network mobility generally occurs on the network-to-network link meaning that one part of the network moves during accessing to some other network. *Session mobility* means that the session moves to the terminal where the user logs in. Session mobility is different in that it does not presume the change on the terminal-network interface but on the user-terminal interface.

III.4.1. Micromobility

Wireless attachment enables uncomplicated and frequent changes of the attachment point. Such network-intern mobility is usually referred to as micromobility. The change of attachment point is typically called inter-cell mobility, as opposed to intra-cell mobility i.e. terminal movements within the geographic coverage area of one access point. The first typically results in a (horizontal) handover meaning that one network access point hands over the mobile terminal to another (similar) access point. Intra-cell handovers are possible, too, usually meaning a resource reallocation at the access point for the moving terminal.

Micromobility is characterized by very frequent local handovers, requiring efficient location tracking and efficient adaptive routing to the terminal. This represents a considerable complexity increase. On the other hand, being limited to one network type, this challenge can be resolved in rather different ways without provoking problems with scalability, compatibility and signaling.

Micromobility can be managed by native network components (as e.g. in GSM) and, in this case, has to be integrated in the wireless technology standard including all necessary protocols and signaling. Alternatively, special mobility supporting entities can be added (on the top or as a replacement of some present components) to existing networks thus requiring external network-independent definitions, typically on higher OSI-layers (like e.g. Cellular IP [34], IDMP [35], HAWAII [36], μMIP / Hierarchical Mobile IP [37] approaches).

III.4.2. Macromobility

As mentioned before, mobility can be provided at higher levels if it is insufficiently supported by the network technology. Especially in the general macromobility case, this seems inevitable either because of the rather different and hence incompatible networks involved or very often simply due to the use of a mobility-unaware backbone interconnecting both networks (today, typically IP-based backbone, e.g. Internet). Though mobility can always be provided at the application level, it results in an additional complexity in every application and remains a specialized solution. Since today IP is the quasi-standard and supported by every network technology, the integration of mobility in IP seems to be a reasonable alternative. Such definitions exist within the IETF standard called Mobile IP [31].

The idea of Mobile IP is to provide mobility independently of the underlying network technology on the lowest common routable layer. Although Mobile IP is a published standard, the interaction of Mobile IP with particular networks has to be studied carefully. Especially the minimization of resulting delays and the possibilities to provide seamless handovers are of great interest, since Mobile IP has been originally designed to provide only hard handovers.

III.4.3. Handovers

The change of the point of attachment of a terminal is called a handover. Handovers are characterized by different criteria, e.g. the administrative nature of the receiving access point (inter-domain/intra-domain), the topological impact of the handover (intra-cell/inter-cell) or complexity of the procedure (planned/unplanned, smooth, fast, etc.)

Inter-domain handovers are handovers between access points of different authorities while *intra-domain handovers* occur between two instances under the same authority. If the change of attachment implies the change of the physical attachment (i.e. change from one base station to another, from access point to the next, etc.), we speak about an *inter-cell handover*. *Intra-cell handovers* only imply changes to the kind of attachment at the same physical instance. *Smooth handovers* minimize the packet loss during the handover while *fast handovers* minimize the delay. *Seamless handovers* are defined as smooth fast handovers, i.e. procedures minimizing both the handover delay and the packet loss. The general aim is to provide a transparent handover preserving established connections.

Seamless handovers are especially difficult to handle in the general case. Such handovers typically require packet buffering in order to bridge the handover delay and either an inter access point protocol or some sort of multicasting since the packets have to arrive at both new and old access points while the handover is being processed.

Seamless handovers are more or less simple to integrate in the networks of high complexity with highly evolved signaling. Indeed, they are naturally supported by all cellular network standards. However, such definitions are completely new in the LAN sector. Modern WLANs typically provide insufficient support for seamless mobility.

III.5. Security Considerations

It is commonly understood that modern telecommunications require security features. Security has become one of the most important considerations in every telecommunications system standardization process.

Here we show why the security in communications is important. We first outline the specific risks. We then give a common overview of modern IT security introducing the typical security functions, the indispensable trust relationships and the security mechanisms.

III.5.1. Risks

The use of every technology contains different risks for the parties involved. These risks need to be minimized to attract users to the technology and to maximize its efficiency. For the communication technologies, such risks can be subdivided into two major groups:

- Infrastructural risks (mainly for the states, enterprises and organizations)
- Personal risks (e.g. for the enterprises, end-users, etc.)

Infrastructure deployment needs considerable economic investments. For a commercial infrastructure (e.g. a telecommunications network), these investments have to be amortized during the infrastructure's lifecycle. Therefore, the commercial infrastructures have to be reliably profitable. Thus, these infrastructures have to be protected since otherwise the anticipated risks outweigh the expected profit: any investment in such an infrastructure becomes too insecure and thus unattractive. For the public infrastructures without profit expectations, a protection is equally indispensable. Although no profit is expected, the infrastructure always has a clearly defined purpose. Any usage beyond this defined purpose must be considered an abuse. If the abuse risk is too high or even outweighs the normal usage expectation, the infrastructure is not viable.

Generally, every deployed telecommunications infrastructure is in danger of being abused, i.e. used in an improper or unwanted way. Such abuse can be of different origins including technical breakdowns and failures, abusive usage by the authorized users and intrusions by the non-authorized users (intentional or unintentional). While the technical failures and breakdowns are part of the safety and technical reliability considerations, the latter two issues can be of a malicious origin thus representing an *attack* on the infrastructure. They are thus subject to security considerations. Such attacks generally account for an unauthorized or incorrect resource usage including namely:

- Unauthorized service access (unauthorized element acting as an infrastructural part or as a user and accessing the service)
- Non-contractual resource usage (authorized user exceeding her contractual quota or accessing out-of-contract services or otherwise changing resource usage information)
- Denial of service (preventing authorized users from accessing the infrastructure, its resources or services)

On the other hand, the usage of a telecommunication infrastructure bears several risks for the user. Generally, a user is concerned about her privacy. In particular, this includes the confidentiality of her data and her location privacy. Another pertinent user concern is a correct billing of the used services/resources (in terms of contract, agreement, laws, etc). Here again, if the risks of abuse are too high, the users will avoid using the infrastructure

ultimately resulting in its complete uselessness. These attacks are generally situated around an out-of-contract information collection on users including:

- Unauthorized access to private communications or data
- Unauthorized location tracking of users
- Unauthorized behavior tracking of users
- Unauthorized actions on behalf of others (imposter)
- Denial of service: preventing a user from using a service
- Incorrect resource usage information

Unfortunately, these risks are not just hypothetic. For financial, social and other reasons, the existing telecommunications infrastructures are a daily target for different types of attacks. Moreover, some international and national laws (human rights, customer protection, etc.) imply or even explicitly require sound measures for user privacy and resource usage transparency for the user. This reality calls for a definition of a *security policy* which identifies the involved subjects and objects, the associated risks, and then prescribes the necessary (typically not sufficient!) protective countermeasures (security functions) and defines their exact objectives.

III.5.2. Security Functions

Today's countermeasures to the presented risks are based on the *security functions* defined in the respective telecommunication standard. Depending on their objectives, security functions can be integrated on different ISO/OSI layers. In a practical communications system, different security functions exist on different layers simultaneously. The classic security functions include:

- Access control and authorization: the access control blocks all requests from unauthorized users. Authorization provides the necessary abstract description of rights for the authorized users (e.g. group membership). Using this description, it is possible to control how the users access the resources (which parameters, which services, etc.)
- Data confidentiality: protects transmissions from sniffing by any third parties
- Data integrity: assures the involved communication parties that the exchanged data is not manipulated during the transmission
- Anonymity, location privacy: avoid user tracking (positional and behavioral). The anonymity function assures that any system identity from which a third party can derive the user's contractual identity is not readable by an untrusted party during any exchanges. Location privacy is a special case of a general anonymity function. Location privacy implies user mobility and provides the anonymity of the mobility-related exchanges.
- Non-repudiation and accounting: provides reliable (or even undeniable) proofs of certain actions, e.g. of resource usage.

III.5.3. Trust

Today, the security functions rely on a pre-established trust relationship (typically provided by an existing service contract or by relationships to a common trusted party) between an abstract user part (user agent) and its abstract counterpart in the accessed infrastructure. The necessary trust relationships between the involved entities constitute the *trust architecture* which can be often seen as a part of the infrastructure management. In the digital world, the trust is typically expressed by a convenient numeric representation: usual trust representations are e.g. passwords (i.e. shared secrets), digital

certificates (i.e. distributed secrets) or/and smart cards (i.e. trusted execution environments containing secrets).

III.5.4. Security Mechanisms

Security mechanisms are the realizations of the security functions. In modern wireless digital communications, the *security mechanisms* are typically based on cryptology and cryptography.

The access control and authorization function are typically based on the *authentication* mechanism enabling verification of the claimed identity. Herein, the identity can be any identifier bound to the service contract in question. In the layered OSI models, the identity is often represented by the characteristic identifier of the layer in question (e.g. the equipment address at L2, the node name at L3, session identifier at L5, etc). The verification of this identity is possible using the pre-established trust relationship. The authentication method defines the necessary cryptographic functions, computations, conditions and the exchanged messages (i.e. the authentication protocol).

Based on this verified identity, the involved parties can make decisions whether to grant access to the demanded resource, to continue the access procedure, etc., thus providing the access control function. The parties can also enforce limitations and guarantees on the established session, thus providing the authorization function.

Data confidentiality is typically provided by an *encryption* mechanism. The encryption can be based on the existing trust relationship. However, today the encryption typically uses temporary secrets agreed upon during the authentication phase (if supported by the authentication method). For further information on encryption methods, see [38].

Data integrity can be provided by message signatures, so-called *message integrity codes* (MIC). MICs are often provided by cryptographic hash functions, parameterized by a session secret.

Location privacy and anonymity can be provided by using a mechanism delivering temporary identities. Such identities can be dynamically agreed upon during the authentication. Another possibility is to hide system identities within cryptographically encrypted tunnels.

The existing mechanisms to implement the non-repudiation function typically rely on personal *cryptographic signatures*. A request and the corresponding signature can be stored in a journal. This journal can serve as a proof in a case of a dispute and can also be used for billing. Note that such signature must not be based on any shared secrets since the non-repudiation function has to provide reliable proofs in a possible contractual conflict case. Another possibility is the usage of *smartcards* i.e. tamper-resistant autonomous devices which independently keep an internal unalterable log of all actions.

It is a common understanding today that the security of all used mechanisms should be based on the secrecy of keys as opposed to the secrecy of algorithms (*Kerckhoffs' Doctrine* [39]). The ensemble of the security mechanisms and the actual secrets agreed upon (statically or dynamically) between the entities to enable secure contact in the sense of the envisaged task define a *security association* between these entities.

Additionally to these cryptographic mechanisms, monitoring mechanisms are often used for the observation of the environment. Such monitoring permits to produce alerts when an unusual behavior is detected and to react to such alerts automatically. Monitoring typically uses network probes deployed in the monitored network. The collected data is sent to processors which compile network usage maps. These data can also be used by

intrusion detection systems (IDS) [40] which alert network administration in case of detected problems (e.g. particular DoS attacks, virus transmission, etc.) Today, the IDS mainly rely on known attack signatures. These are thus comparable to the virus scanners: if the attack is not known, it can not be reliably detected. The signature database must be kept up-to-date. IDS capable of detecting "anomalies" i.e. unwanted situations based on behavior analysis are a current research topic. Once the attack has been identified, dynamic filtering mechanism can be used to stop the ongoing attack and to prevent further problems (*intrusion prevention systems*). Such filtering is typically based on *firewalls* [41], i.e. rule-based access controllers.

Among the risks listed above, both users and infrastructures are concerned about denial of service (DoS) attacks. However, theoretically it is impossible to completely avoid this kind of attacks. This is because certain normal system conditions (e.g. too high system load) can result in a practical denial of service for certain users. This attack is thus theoretically indistinguishable from routine situations. Judging what is an attack and what is not becomes difficult.

The practical approaches against this kind of attacks include *service presence hiding* and *deterrence* techniques. For the telecommunication systems, service presence hiding equals transmission channel hiding. It is mainly used in military telecommunications systems. Deterrence techniques include rendering the attacks too complicated, too risky or simply uninteresting for the attacker. In public systems both techniques are generally not applicable: a commercial service typically relies on massive advertisement of its presence, its technical characteristics and tries to provide an easy and comfortable access. Compatible equipment is supposed to be available on the mass market.

Current countermeasures against specific DoS-attacks exist and include adaptive filtering [42]. These build on monitoring mechanisms with decision systems detecting the attack and filtering out attack-specific traffic.

III.5.5. Wireless Security Challenges

Wireless communications impose tighter constraints on security [43][44]. Indeed, using wireless communications we lose the supposed privacy of the wire. This has two major implications.

Firstly, with wireless links there are no implicit guarantees on identities of receivers and transmitters. While wires start at some well-known position leading to another well-defined position else over a fixed path, wireless transmissions typically propagate randomly depending on the environment (see Sections III.3.2 and III.3.3). Thus, an explicit and reliable identification of all concerned authorized entities is indispensable.

Secondly, in the most general case, it is both difficult to determine from where the signal came (exact positioning) and to predict who will be able to receive it (propagation ambiguity, III.3.3). Thus, a wireless transmission raises serious concerns about data integrity (constant uncertainty about the data source) and major concerns about data confidentiality (the data are typically received in the whole coverage area). Given the unpredictable character of the propagation, prudence is appropriate in all assumptions about possibilities to receive, modify and decode the emitted waves even with highly distance-limited or directional transmission technologies. This is demonstrated e.g. by the network imaging project in [45].

The security issues discussed until this point are directly implied by the changed physical characteristics of the medium: wireless is different because of the wire absence. However, for the same reason, wireless technologies also provide an increased flexibility promoting

or motivating new services like e.g. terminal mobility or localization. Yet, every new service can be abused and thus bears a risk. Thus every new service is subject to new security considerations.

Terminal mobility can radically change the security considerations. A mobile terminal can potentially visit different domains and authorities. It thus connects to and over foreign networks which are not necessarily under the responsibility of its own authority. This already implies that a security policy has to additionally consider this new possibility and define rules where and how the mobile node can and should connect to and where it cannot or should not. However, the implied identification requires scalable trust mechanisms since a manual a priori trust establishment is difficult to combine with a free mobility definition. The other category of risks comprises viruses, worms and, generally, every software installed during the absence of the mobile (i.e. not by the responsible authority). This software has to be compliant with the security policy of the operating authority which is not trivial to achieve. The mobile itself thus becomes a weak element. Other mobility-related risks include impersonation by an attacker during the absence of the authorized mobile, diverse new DoS attacks exploiting the mobility management mechanisms, attraction of mobiles towards attackers, malicious handling of mobile traffic by a serving (foreign) network, etc.

Terminal mobility also implies new service needs. Terminal/user localization is a service enabling location based user services (LBS) for mobile users. Localization also adds its own risks: if the localization service is abused by an attacker, the location privacy of the user is violated.

From the user's point of view, wireless technologies provide new possibilities. With new possibilities come new risks. Additional protection mechanisms are required, adding to the system complexity.

From the attacker's point of view, wireless technologies are an interesting target. Indeed, current wireless technologies can provide broadband Internet access or serve as access networks to core enterprise networks. Moreover, wireless technologies offer new attack scenarios (e.g. user localization, new DoS attacks, etc.) and even new attack paradigms (e.g. the so-called *wardriving*). At the same time, the associated risk for the attacker (e.g. the risks being detected or caught, etc.) is actually smaller than with the comparable attacks in the wired world. Most importantly, attacks against wireless technologies often do not require the direct physical presence of the attacker. The determination of the attacker's position can vary from very difficult and unreliable (because of the discussed physical phenomena) to impossible (pure passive data sniffing attacks).

The last two points constitute what we call the *wireless security dilemma*: wireless technologies are more difficult to protect but easier to attack.

Because of this situation inherent to wireless communications, a certain set of necessary security functions has to be supported by the technology in question. It is typically problematic to try to add all of the security functions in the higher layers since this can not exactly correspond to the risk associated to the wireless technology itself.

Hence, wireless technologies need to define and to implement security mechanisms. These require implementation prudence. Security implementations are complex and error-prone. They increase computational resource requirements and should provide for flexible update possibilities. Moreover, the existing security mechanisms rely on pre-established trust relationships. The trust establishment requires additional management efforts and management functions. Thus, the security considerations also increase the complexity of the wireless systems.

III.6. Negative Synergy Effects

The problems described so far are presented as stand-alone problems. However, as has been pointed out, some of the stated features (e.g. security) are indispensable for wireless technologies and others (e.g. mobility) much sought after.

Unfortunately, in practice the situation gets worse when the goal is to build a wireless network providing two of the above features simultaneously. It is important to notice that the aggravation is usually greater than the "sum" of both base problems. That is what we call a *negative synergy effect*. In what follows, we demonstrate this effect with the examples of two arbitrary feature combinations. However, this effect is not limited to the mentioned combinations and also occurs in other cases (e.g. QoS and security).

III.6.1. Security of Mobility

Let us assume that a wireless infrastructure network is to provide a secure mobile access to its users. We also assume that we have both a solution providing a required security level and a solution providing mobility functions. Now, the negative synergy effect manifests itself by the observation that it is typically not enough to combine the two solutions to try to achieve the goal.

Typically, both the security and the mobility solutions have to be adapted or reengineered. For instance, the security concept has to be prepared for possible roaming, i.e. to handle access by users from foreign networks (of the same type). A scalable inter-domain trust architecture has to be designed. The mobility subsystem has to be integrated in this trust architecture. Its performance has to be reevaluated in the light of authentication procedures taking place during e.g. a handover to a new access device. On the other hand, we have to extend the inter-device and inter-network mobility protocols (handover mechanisms, etc.) to carry the authentication data.

Generally, this combination results in a higher complexity of the protocols and the network architecture.

III.6.2. QoS and Mobility

The dynamics of the wireless links represent serious problems in guaranteeing quality of services (QoS). Combined with mobility, additional problems arise like delays (in handover case), resource management (are sufficient resources available at the new attachment point?), re-negotiation and re-reservation (new attachment point can not provide negotiated link parameters), etc. So, mobility protocols have to be extended to support different service classes. Additional exchanges may be necessary to probe available resources. The mobility manager has to be prepared to understand additional QoS attributes and the QoS manager has to be changed in order to react to mobility events. Similarly, these features can not co-exist independently.

III.7. Conclusion

The proliferation of the Internet and the increased business demand for ubiquitous data access require flexible implementations of anytime, anywhere, always-on concept. In this chapter we identify the access network problem currently constituting the major technological barrier for such global provision of broadband data access.

How can we resolve the access network problem? Currently, the so-called 2.5G systems – either following the GRPS/EDGE [19] or the competing cdma2000-1X [20] standards – are being deployed virtually in all countries. Hence, although the new broadband services suggest higher data rate support on the user link than what is provided by 2.5G, the wireless technologies have proven to be reliable enough for an economically reasonable world-wide data network deployment.

In this chapter we thus classify the available wireless technologies. We analyze the difficulties with wireless in general and state that wireless represents a technological challenge. Generally, no wireless technology is currently capable of fulfilling all user wishes and tradeoffs need to be taken into account. That is true in terms of the special dependency illustrated in Figure III-3 (mobility vs. data rate) but also for a lot of other reasonable combinations (provided service quality vs. data rate, price vs. security, network management vs. complexity, etc).

Yet, taking a look at the technological development of digital wireless communications in the last years, we observe that the wireless communication sector is one of the most rapidly evolving IT sectors. This fact and the proven suitability of wireless for world-wide deployment provide a motivation base for treating the access network problem as a technological rather than an economic problem. If we could provide users with an easy and efficient wireless broadband access to data services without new cost explosions for the operators, the concept would be likely to work out.

In the next chapter we take a closer look at the suitability of particular technologies for specific services and try to understand how the future wireless system should look like.

C H A P T E R I V

4G Wireless Systems

In this chapter we give arguments for the transition to the new fourth generation of mobile telecommunications systems (4G). Based on the experiences with previous wireless systems, we try to provide a time frame predicting the future 4G development.

Based on this motivation, we present state of the art issues in the current 4G research. We first introduce several anticipated issues with 4G terminals. We then show the different possible approaches to 4G and synthesize the related work, summarizing different alternative 4G visions.

We then present the anticipated technological challenges implied by the European Commission's 4G vision. We explicitly discuss such problems as security of heterogeneous systems, mobility in 4G, heterogeneity of signaling and the heterogeneity of the provided services.

We conclude this chapter underlining the 4G challenges on which this thesis focuses.

IV.1. Motivation for a New Generation

Our argumentation for the next fourth generation of mobile systems (4G) is separated in two major parts.

The first part gives argumentation for a new generation as such, making assumptions on future telecommunications landscape on one hand and taking into account the encountered problems and weaknesses of the current generation on the other. The futuristic vision helps to identify features, services, etc. that may become necessary in the future. By projecting the current generation into this futuristic landscape we try to understand to which degree the current generation would fail to fulfill the anticipated features and discuss how the next generation should be designed. With this approach we can thus establish some basic requirements on the next generation.

After the discussion of the general need of the next generation, the second part identifies the timing constraints for the transition. Analyzing previous generations (1G, 2G) we draw analogies to the current generation in terms of conception duration, conception to market delays and system lifetime. We also discuss the uncertainty associated with this approach and the possible additional delay factors.

With this motivation, we then discuss some expectations of the future generation.

IV.1.1. Substantial Argumentation

In this section, we discuss the suitability of different candidate technologies to bridge the access network problem presented in the previous chapter.

Third Generation of Mobiles

The third generation of mobiles (3G) was expected to be the future global standard for the integrated voice and data communications. 3G was designed in the last decade of the 20th century with the goal to provide enhanced wide-range voice and data services. Principally, 3G aimed at the improvement of the radio link performance in the 2G scope. Although the developed standard features drastically improved data rates as compared to 2G, from the point of view of the data services the practically available data rates can be still considered scarce. This can be directly observed in a direct comparison to the development of the wired technologies providing home Internet access. In the last decade (i.e. from 1994 on), the phone-line Internet access technologies have evolved from V.34 modems (28.8kbps) over V.90 (56kbps) to cable (1-2Mbps shared) and ADSL (originally 500kbps, today up to 10Mbps). This means an almost 350fold increase. At the same time, the data rate of the wireless cellular access is lower and has not been able to keep up the pace. From the original GSM CSD service introduced in 1994 and providing 9.6kbps, the cellular systems have evolved over GPRS (about 64kbps in practice) to EDGE/cdma2000 RTT-1X (typically about 100-130kbps). The emerging 3G (e.g. UMTS) provides about 300kbps in practice. This corresponds to a 30-50fold increase since 1994. Moreover, the provided data rates highly depend on the network operator's overall capacity, the number of users in the cell and the distance to the base station.

However, the relatively limited data rate is not the only problem of the 3G data service. Because of the vast, national-scope infrastructure and many intermediate nodes, the user experiences a high network latency (e.g. from the point of view of IP, the whole 3G infrastructure is one link). GRPS and EDGE exhibit network *round trip times* (RTT) of 600ms and more. UMTS links are expected to be better, but they still have RTT of about 200-250ms. The planned W-CDMA/HSDPA service (High Speed Downlink Packet Access), defined in the 5th release is expected to have more than 100ms RTT, i.e. almost an Internet-level RTT. Such high network latencies are inappropriate for certain application classes: interactive applications require less than 400ms overall RTT to remain efficient (the end system is not necessarily within the UMTS backbone and thus the typical Internet RTT of about 100-200ms has to be added to the 3G latency); VoIP

and similar applications (videoconferencing, etc.) require a round trip time (RTT) to be under 250ms; in some existing popular interactive online games (e.g. id Software's Quake, etc.), the maximal acceptable RTT to the game server is required to be under 100ms in order to provide a fair chance of winning and a good game experience.

Furthermore, being a fully managed commercial mobile network, 3G targets national scope telecommunications providers. Like 2G, 3G uses a license model to prevent random medium access by non-authorized organizations. Since the licenses are expensive [28], this implies a major telecom operator with a mammoth infrastructure behind every 3G RAN. To fulfill the requirements of this authority, the 3G RANs are designed to be reliable and manageable and to support different qualities of service. This justifies the high cost of the 3G equipment. At the same time, this limits the competition on the market to few license holders who not only have invested a lot in the infrastructure but also have paid a high price for the license. The operators have to amortize this fixed cost and the current variable maintenance and management cost over the user services provided by the infrastructure. Thus, 3G RAN access is likely to remain costly. It is unclear if attractive flat-line pricing models are applicable to such infrastructures. Current per byte or per minute pricing is rapidly revealed unsuitable for the always-on paradigm.

Hence, a consequent national-scope investment is needed for 3G advantages to materialize. This is however difficult to afford, especially in developing countries where big investments are particularly risky. In a focused coverage, 3G comes at a very high cost per bit compared to other, more data-centric technologies like local or metropolitan area networks. That is one of the reasons why the 3G systems had a difficult start. They are primarily being deployed in Japan, South Korea, Taiwan, Hong Kong, Indonesia, a few countries of South America, Australia, New Zealand, western Europe and North America [19][20]. Figure IV-1 [19] summarizes all actual and planned commercial launches of the 3G system from the European point of view (W-CDMA/UMTS). It shows that the developed countries prevail.

Although the slow 2G-3G transition process has now started, so far the 3G systems do not seem suitable to provide a broadband data service deployment. In the developed world, these are often considered technologically inadequate. For the developing world, the technology needs major investments.

Personal Satellite-based Communications

Personal satellite-based communications that do not require any local infrastructure initially seemed an interesting alternative technology to provide global data coverage. However, these links do not provide sufficient data rates. Also, with the commercial difficulties of the global satellite phone network projects (such as GLOBALSTAR [46], IRIDIUM [47], etc.) in the last decade of the 20th century this hope has mostly faded away.

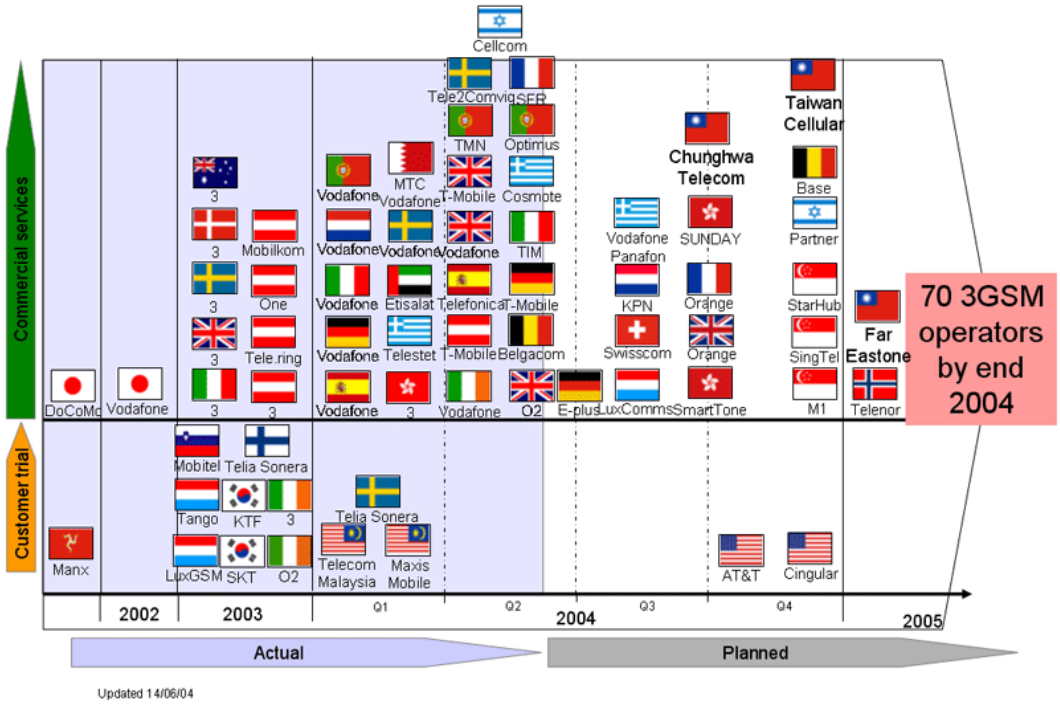


Figure IV-1 3G: current and planned UMTS launches

Alternative Technologies

The developed world still suffers the consequences of the deep IT industry crisis that followed the Internet-euphoria in the beginning of 2000. The additional investments in the Year 2000 (Y2K) bug fixes and the following 3G spectrum license shock in Europe [28] have dug major liquidity holes in the balances of the whole industry branch. Moreover, it has killed the investment intentions in this sector in the upcoming years. That situation still accounts for 3G deployment hesitation. The mobile broadband data access so far remains an unfulfilled wish.

Alternative wireless technologies and their integration into the core networks with new technological paradigms in mind could bring back the focus to where it should be – on the user – and the investor confidence to the industry. They could complement the 3G in the technological and economic sense or culminate in the definition of a new global post 3G standard.

The existing alternative technologies primarily include wireless metropolitan and local access networks (WMAN/WLAN). In contrast to 2G and 3G, these networks explicitly aim shorter ranges like e.g. a city area, an enterprise, home networking, etc. Standardization institutes in Europe (ETSI) and in the US (IEEE) recognized the need for such solutions quite early (ETSI: mid-1991). ETSI BRAN [17] had accomplished the necessary preliminary work by 1993 and later published its WLAN (HIPERLAN/1, 1996; HIPERLAN/2 [15], 1999) and WMAN (HIPERMAN, HIPERACCESS) standards. Although ETSI started earlier, the corresponding IEEE standards (IEEE 802.11 [14], IEEE 802.16 [16]) were available sooner and had more market success. In the recent years, a collaboration between ETSI and IEEE (e.g. in 802.11a-HIPERLAN/2, 802.16a-HIPERMAN) ties together the common standardization efforts.

IEEE 802.16 (also called WiMax [18]) is a WMAN standard that aims to provide fixed wireless broadband access and thus addresses the wireless local loop (WLL) and the last

mile. Using millimeter waves, 802.16 uses directional wireless links (line of sight) explicitly aiming at the served destinations. The base standard is thus not meant to provide a radio coverage but to serve as a backbone for other network installations. Currently, it can not be used by the end users. However, new standardization efforts (802.16e) are being made to provide area coverage and limited mobility support.

IEEE 802.11 [14] typically aims at the SOHO market (small offices, home). 802.11 provides a wireless local area network interface. 802.11 features different physical specifications unified by a common logical link control. Thus, from the user's point of view, the usage of any 802.11 transmission technology is very similar to that of a typical wired network today, like the very popular Ethernet. As a LAN, 802.11 features very short delays (under 1ms) and high data rates (up to 54Mbps with 802.11a and 802.11g). 802.11 WLANs do not need any spectrum licenses and can thus be freely operated by local WISPs, businesses and even private users. The available WLANs can not provide the same spectrum efficiency as the 3G and are thus not more economical in the global coverage. However, WLANs are much cheaper for a focused coverage.

Thus, 802.11 WLANs provide a true LAN experience at a low price. Yet, the problems of the alternative technologies lie elsewhere. Today, the main issues with the 802.11 WLAN technology are infrastructure security, user quality of services and network device management. The base standard does not allow for a comprehensive user quality of service support. Since 802.11 does not require any spectrum license, the regulation authorities limit the allowed maximum emission power levels. The coverage area of a 802.11 access point is thus also limited: a usual omnidirectional antenna provides roughly 50m indoor and 300m outdoor coverage[1]. The limited cell size requires multiple access points to be deployed. That turns out to be difficult: as a "wireless Ethernet", 802.11 hardly defines any high level management functions.

These interdependent issues can be resolved through consistent and rapid management mechanisms. However, the original standards do not define such mechanisms. Thus, these have to be added on top of the existing network, which is sometimes not possible and otherwise often results in inefficient compromises and proprietary solutions.

Another issue which appears only insufficiently covered is the mobility support. 802.11 features link layer mobility support: the user traffic is switched to the new access point automatically as long as the used access points are connected to the same Distributing System (typically a wired Ethernet switch hierarchy). This works quite well in a small unmanaged network but there are some scalability and security concerns in bigger installations.

Thus, the IEEE 802.11 can be seen as an interesting data transport technology but it lacks a comprehensive network management support needed for global deployment.

A New Global Standard

An apparent alternative is to conceive a completely new, global wireless standard. Such a revolutionary approach could provide the ultimate technological answer. Indeed, a new working group has been founded at the IEEE with the purpose of conceiving a new wireless cellular system. Called 802.20 [21], this group explicitly aims at providing users with low latency / high data rate wireless access thus featuring a local network experience over a global network. The targeted capability of providing low latency links and the explicit requirement for the support of different service classes would make this technology suitable also for voice and video calls, multimedia data real-time transport,

[1] The exact distance varies a lot depending on miscellaneous physical properties, see III.3.3.

etc. However, by now, the 802.20 WG [21] has only finished its requirements framework. Hence, the associated uncertainty is unacceptably high. Above all, such a new technology would have to provide answers to the obvious question of how it aims to be more successful than the already existing technologies, i.e. 3G. The associated deployment delays are likely to be high due to the necessary conception, test and evaluation phases. The performance under field conditions and the scaling of such solutions must be carefully studied.

Opportunistic Integration of Different Technologies

In the mean time, we could think about possible integrations of the existing technologies, explicitly including both 2G and 3G. This approach exhibits diverse appealing properties, such as deployed infrastructure preservation, economic advantages, flexible focused deployment, a faster availability, etc. It is thus capable of reducing the associated investment risks from the point of view of the operator, resulting in reduced costs and sometimes better performance for the users. However, this approach also accounts for an increased complexity related to the increased heterogeneity of the overall system. The associated technological challenges require new design paradigms and represent the main goal of this work.

Synthesis

The discussion above is summarized below in Table IV-1. It evaluates the suitability of the different approaches to the access network problem in terms of their availability, cost and various technological parameters. Obviously, the values for the new wireless standard are estimated based on the conviction that this could be designed to optimally fulfill all technological criteria.

As can be seen, the integration of technologies has several interesting properties generally providing the flexibility in terms of almost every criterion. However, such integration is a technological challenge because of the necessary heterogeneous system design.

Table IV-1 Different approaches to the access network problem

	3G	Satellites	Alternative WLAN, WMAN	New standard	Integration of different technologies
Availability	Available	Available	Partially available	Not available	Partially available
Focused deployment cost	High	High	Affordable in a local scope	High	Flexible
Data service	Restricted, expensive, high latency	Too low data rates, high latency	High data rates, low latency	High data rates, low latency	Depending on the used access network
QoS support	Integrated	Integrated	Insufficient or underspecified	Integrated	Yes for voice, limited for data
Network management	Integrated	Integrated	To be done	Integrated	Difficult

IV.1.2. Timeline Argumentation

One could argue that the future-oriented argumentation given in IV.1.1 is too speculative and not convincing. Indeed, at the moment, when the commercial 3G networks are being deployed in the European centers, it is quite difficult to claim to have a clear vision with regards to the needs of the technology generation which might follow the one currently

deployed. This is especially true for such rapidly developing industry sectors as telecommunications and electronics[4].

Moreover, the previously given argumentation is not suited to set any time constraints. Time is however one of the most important issues in this motivation: for the sake of the general human progress, everybody would spontaneously agree that 4G will exist one day. The real question however remains why the community started thinking about 4G around 1999-2000 and why we continue discussing the identified problems now.

In this section we will thus give a different argument for 4G that not only is free from any futuristic vision of technological needs but also outlines delays until 4G might become true. For this purpose, we want to recall some facts of the past.

The History of Mobile Networks

Table IV-2 summarizes the history of the public land mobile networks (PLMN) development from the European point of view as presented in [49]. In particular, it illustrates the repeating approximate 10-year cycles both in the conception phases and in the generation lifetimes.

Table IV-2 10 years cycles in the mobile networks (from a European view)

Year	Milestone	Cycles		
1981	*Commercial deployment of NMT:* ***1G start***	1G to 2G: 10 years		
1982	Creation of Groupe Spécial Mobile (GSM) at CEPT[5]			
1984	Commercial deployment of AMPS networks in the US			
1986	Big number of users leads to NMT extensions			
1988	Big number of users leads to AMPS extensions			
1989	European Union RACE Project "invents" UMTS		3G conception: 10 years	
1992	World Administrative Radio Conference (today: WRC) allocates 230 MHz to Future Public Land Mobile Telecommunication System (FPLMTS).			
1992	*Commercial deployment of GSM:* ***2G start***			2G to 3G: 10 years
1994	Second wave of UMTS research projects			
1995	RACE vision of UMTS			
1996	Creation of UMTS task force			
1996	*Digital overcomes analog*			
1997	Establishment of the UMTS Forum			
1999	UMTS decision			
2000	WRC designates IMT-2000 extension bands			
2002	*Commercial deployment of UMTS:* ***3G start***			

The variety of incompatible networks and the increasing popularity of data services have motivated and much influenced the work on 3G. In 1992, i.e. at the same time as the commercial deployment of first 2G networks had started, the ITU allocated frequency ranges for the next generation of PLMN (then called FPLMTS) thus providing an international common base for 3G. Finally, in 2002 the first commercial 3G networks were commercially deployed in Japan.

[4] But, as Jules Henri Poincaré remarks in "Science and hypothesis" [48], "it is far better to foresee even without certainty than not to foresee at all".

[5] CEPT, i.e. Conférence Européenne des Administrations des Postes et des Télécommunications, is the creator and standard-body predecessor of today's European Telecommunications Standard Institute (ETSI)

The Anticipated 3G to 4G Transition

In regards to 3G, the observed 10 years cycles seem to continue. The first research concepts aiming at 3G appeared about 1989. The spectrum was reserved by ITU-R's World Radiocommunication Conference (WRC) [50] in 1992, i.e. at the same time as the first 2G networks were deployed. The active technological development of 3G started with the creation of the UMTS task force in 1996 and culminated in the UMTS decision in 1999. The largest parts of the standards were accomplished by then.

Consequently, the first projects naming 4G started in 1999 and the first dedicated thoughts about B3G and 4G systems appeared in the international research press about 2000-2001 [49][51][52][53][54], i.e. just before the first commercial 3G networks were deployed in Japan. In 2000, the World Radio Conference allocated 3G extension bands which were to be used in the B3G scope. All this corresponds to the 10 year cycles illustrated in Table IV-2.

Continuing along this line, the concrete shapes of 4G should be clarified by the end of 2005 and the active technological 4G development should start about 2006-2007. This should be roughly finished by 2009, with several detail issues being addressed in the following years. The first commercial systems could then be operational by 2012. However, this presumes that no additional delays occur.

Possible Delays

At least in Europe and in the US, the 3G deployment currently seems to be delayed. Indeed, by the end of 2004, not all western European countries have started the 3G deployment (see e.g. Figure IV-1). Also, the deployment process is starting quite slowly, being often limited to some few centers. The critics of 3G claim that the reasons for this could be in the developed technology itself. Indeed, one could argue that 3G (in Europe: UMTS) is too complicated and too costly to become successful. One could also criticize the fact that the original goal of creating one common global standard has not been achieved since different concurrent versions of 3G are being standardized and deployed, in some extreme cases within the same country (e.g. Japan has deployed both cdma2000 and W-CDMA). However, the deployment of the alternative technologies (like e.g. 802.11 hotspots or WiMax) also lags behind the expectations that have predicted a WiFi-boom and hotspot number explosions by 2005 – which so far have failed to become true. There is no doubt about the popularity of WiFi. However it is not booming, it is being carefully developed. The real reasons thus could be either of a social (e.g. a simple current disinterest in mobile data) or of an economic nature (too costly in deployment, too risky for operators; too costly, too complicated for users, etc.).

We tend to think that economic barriers prevail. Indeed, businesses have so far often expressed their need for mobile communications development. This has been much discussed in different business scopes: home- and telework, instant data access for mobile sales personnel, fleet management, reduction of infrastructural costs, globalization, etc. With further development of the Internet and associated technologies, private users are also likely to be interested in services such as mobile e-commerce, online gaming, private communications (e.g. voice or instant messaging), various personal and business data exchanges, etc.

The telecommunication crisis initiated by the complete flop of the exaggerated initial Internet business activities (often referred to as the *bursting of the Internet bubble*) seems to be one of the key economic factors responsible for the observed 3G deployment delays. Indeed, the investments in the IT and telecommunication sectors have since radically switched from headlong promiscuity to skeptical cautiousness. From the European point

of view, the starting crisis was amplified by the UMTS license auctions in 2000-2001 raising cumulatively over 100 billion USD in the Western European countries [28].

The paid spectrum prices washed away much of the liquidity of the Western-European telecommunications operators. Yet, this liquidity was necessary for the deployment of the network (infrastructure updates and add-ons). Since the UMTS can not substantially improve the GSM voice service as such, the only added value of the UMTS are the improved data services. Hence, compared to classic GSM offerings, the paid auction price for the UMTS licenses must be amortized over time over the new services which UMTS is about to propose. However, this could render these new services particular expensive.

IV.1.3. 4G Expectations

With ongoing globalization, world-wide communications become an essential service. The 3G, meant to provide a global communications standard, has mostly failed to do so. Instead it now uses different standards in different countries. Moreover, 3G remains a closed "big company" telecommunications forum. That results in the situation where users still need costlier multi-band, multi-technology handsets yet can not access the 3G services using other devices over newer radio access networks. To provide users with a world-wide service we need open flexible standards, also suitable for the Internet and data communications deployment in the developing countries.

At the other end, personal communications are being rapidly developed using short range radios. These need to be considered for the next generation communications because their rapid development is a fact (see Section III.2.2). The existing PAN and LAN technologies are often used for device-to-device data transfers but can easily do more than that. Wireless headphones, handsets and PDAs can already build personal networks capable of data and voice transport. In the home area or in vehicles (e.g. personal cars), this can be extended to home-wide LAN communications. The aim here is to give users access to their data independently of the device currently in use. So, handsets can be asked to dial numbers stored in the home PC and to direct the voice flow to the wireless headphones. Wireless sensors are already available e.g. for outdoor weather condition measurements. Wireless sensors are more and more used in cars. They are also expected to be further developed for home users (intelligent home). This underlines the increasing part of the machine-to-machine (M2M) and network-to-network (N2N) communications in the future communications landscape.

The obviously challenging scenario is to provide users with a bidirectional communications possibility to their personal Intranets independently of their location (anywhere), thus combining the two topics discussed above. These WAN/MAN/LAN/PAN spanning communication sessions have to be secure, reliable and economically reasonable. Also, communications become ubiquitous. The used technology needs to be able to reply to this challenge, providing the best available connection anytime, any place. Existing standards do not allow for this usage.

However, the existing technologies are not competitive. They are more and more understood as complementary. Indeed, the WLANs can easily provide a true LAN experience in limited areas at a low cost while 3Gs RANs can provide true mobility, quality of services and vast coverage. The idea to try to integrate both technologies is thus straightforward.

Taking into account the previously observed cycles and the current delays, we could try to compile a prognosis on the B3G and 4G development in the next decade. The current situation and our forecast are illustrated in the Table IV-3.

Because of the true need for mobile broadband data access, we believe that 3G will be eventually deployed in the business centers of the developed countries despite the currently observed delays. In Europe, this process could be further promoted by governmental policy planning to partly reimburse some license fees. However, the delays and the high license fee [28] have already motivated the development of and the investments in the alternative transmission technologies e.g. IEEE 802.11.

This development, if commercially successful, will lead to a situation with several parallel infrastructures installed in the European centers by 2007-2008. While the 3G infrastructures will be homogeneous, they are likely to remain more expensive. The alternative offerings will be cheaper but are not likely to provide the same service quality. Because of the required spectrum licenses, the 3G systems will be owned by the same national-scope operators. The alternative technologies are license-free and thus enable a free network deployment. These can be owned by both global big telcos and small local WISPs.

The convergence between the different infrastructures will start because of the economic and technological limits of the used technologies. The big operators will try to reduce their service cost by integrating the alternative transmission technologies such as radio access networks (RAN) into their 3G infrastructure. However, this integration will still be much more complicated and costly than a new deployment possible for a small WISP. At the same time, the small WISPs will encounter increasing management problems with the growing user basis and the user traffic. It will hardly be reasonable to add a 3G infrastructure upon the existing one as the control plane. Given the lack of standardized methods, the alternative infrastructures are thus likely to be managed in a proprietary way requiring specific access methods. This will produce the demand for standardization.

Users will buy newer products equipped with further wireless technologies. Deploying these products at home, users will be interested in accessing the combined data. Different devices will be capable of several access methods. (e.g. a wireless ADSL router). Users will be incited to open their hotspots for the usage by the others. For instance, the major French telecom provider plans to propose a reimbursement plan for its ADSL users if they provide WLAN access to its cellular customers over such devices.

Table IV-3 Possible 3G development in the next years

Year	Milestone	Cycles
2003	*European 3G start*	3G to 4G: 10 years
Until 2005	Different 4G visions and 4G research projects	
2006	3G deployment in all business areas in the developed world	
2006	Broad deployment of alternative technologies (WiMax, WiFi, etc.)	
2007	Further deployment, different UMTS updates (HSDPA, HSUPA) and integration of alternative technologies in the UMTS infrastructure	
2008	Convergence of different 4G views implied by the economic and technological factors	
2009	The high popularity of data services shows 3G transport limits and WiMax/WiFi management limits (security, mobility, usability, etc.)	
2010	Deployment of first B3G (3.5G) systems	
2011	Establishment of a 4G forum	
2012	Mature technical drafts of 4G systems integrating different technologies	
2014	*First commercial 4G services*	

Meanwhile, the research will push towards unified and concrete B3G and 4G views. To protect the investments, the deployed alternative infrastructures are likely to be given the necessary attention in this development process. The result will likely be a system providing for a convergence between the different technologies.

While the new 4G architecture is conceived and matures technologically, 3.5G systems are likely to appear on the market by 2010 at latest, filling the gap between broadband and manageable. These updates of the radio link and of the backbone infrastructure could provide the basis for the later expected 4G much in the same manner as GPRS (2.5G) has required and accomplished the necessary infrastructural changes for the transition process from 2G to 3G. The commercial and technological convergence and the available B3G systems will provide the drivers for the establishment of an industry group (e.g. 4G forum) that will be given the task of 4G system standard development. Based on the situation and the previously accomplished research, it could produce mature system drafts by 2012 and the first commercial 4G deployment could take place about 2014.

IV.2. 4G Terminals

With respect to the expectations on the future generation wireless networks presented above, the 4G terminals are expected to have multiple physical access interfaces (multi-mode terminals) with the ability of simultaneous read/write access to several interfaces. Hence, the terminal has to be additionally responsible for security, location and mobility management, QoS channel reservation and dynamic reconfiguration over multiple incompatible links. However, the software complexity of the terminals has to be discussed in respect to the decreasing device size and scarce resources.

In the following sections, we discuss the resource limitations in three different dimensions: computational, battery and input/output system (I/O).

IV.2.1. Computational limitations

Basically, every usual desktop workstation or server could be connected over a wireless adapter. However, generally, wireless terminals are portable devices i.e. smaller, battery-driven appliances. Existing laptops and notebooks are the most powerful portable machines in terms of available CPU and memory resources, output and input devices, communication interfaces, etc. Yet, compared to the desktop workstations these are still limited. Sub-notebooks, PDAs and smartphones represent the gradual decrease of both the size and the available computational resources and battery life.

The constrained capacities (CPU, memory, bus, etc.) limit the possibilities of providing the strongest available security solutions on mobile devices resulting in installation of either new or proprietary security measures for the wireless communications [55]. However, both approaches are questionable in terms of security. A proper reuse of the well-known security mechanisms is considered a more prudent approach.

The popular personal digital assistants (PDA) have started their gradual migration to a so-called personal mobile assistant (PMA), i.e. a PDA disposing of a set of network interfaces. The PMA is sometimes believed to be the killer-application of the next generation mobile networking since it can provide an all-on-one solution for business communications bringing together the voice and data communications like automatic synchronizations with the enterprise's information system.

In this sense, the PMA is a superset of a PDA additionally featuring a comprehensive set of communicative applications. This underlines the hard constraints on the required resources (wireless networking, intelligence for new service support, management, security, etc.) vs. power budget (i.e. battery capacity). However, it also reveals new challenges concerning available human interface (machine-human, M2H).

IV.2.2. Battery Problems

Being portable, the wireless terminals typically make use of a battery. However, whereas the technological progress in both communication speeds and CPU performance follows the exponential development according to Moore's law [24], the progress in the battery research is linear [44]. In that context, the use of complex, resource-greedy applications is problematic. However, some of these (e.g. security, management logics, etc.) might be indispensable. Strong security measures significantly degrade battery life due to the associated computational complexity. This results in a so-called *battery gap* in general and in the *wireless security processing gap* in particular [44].

Today, a typical wireless phone already features 2-3 communication standards, a camera, a color screen, a loudspeaker, an audio player or FM radio, etc. Additionally, user expectations on wireless devices grow with the increasing competition on the market. As opposed to the current use, next generation wireless devices will have to share the limited energy budget over several wireless interfaces which can be used at the same time. The management of these interfaces have to be designed with great care.

IV.2.3. M2H Interface

The use of the most popular data applications like e.g. Web or email is already constrained on the existing PDAs because of screen size, screen resolution, limited and cumbersome input possibilities, etc. Next generation devices should be small, light-weight and easy to use. However, they are also supposed to be fully communicative devices supporting several different wireless standards and a complicated logics for the management of the connections/sessions/etc. Thus, PMAs are likely to tighten the existing constraints since their implementation and the provided features are more diversified and thus more complex. It is not quite clear how this challenge can be resolved.

IV.2.4. Synthesis

To enable the development of such next generation terminals, different efforts are being currently made. One possible approach is reducing the overall design complexity by using a reasonable degree of homogenization. This can be done e.g. by the introduction of generic open abstraction APIs. New ISO/OSI sub-layers have been discussed in some publications [56]. Middleware paradigms [57] basically head in the same direction proposing standardized high-level service access functions.

Lightweight subarchitectures are being proposed for different management issues like the security management architecture in [58]. The authors describe security architecture for 3G/4G devices. According to the authors, the security mechanism has to be lightweight, reconfigurable and capable of capturing the dynamism and agility of mobile environments. Using only well-known security mechanisms like Kerberos and SESAME, the authors want to provide the main security goals. The authors provide a Kerberos-compatible Java implementation of a component based security mechanism that can be added in 3G and even 4G. At the terminal side, the client could be implemented in the terminal itself or in the SIM card. The proposed solution provides Kerberos services with different symmetric, asymmetric and hash protocols, privacy using asymmetric cryptography, inter-domain authentication, role management and key management (enhanced Kerberos).

Simultaneously, the research on integrated circuits, embedded computing, etc. push forward the hard- and software capabilities of new terminals. Also, existing software bricks are being optimized to avoid new divergence [55].

IV.3.4G Approaches

Since the future is difficult to predict, the definition of 4G is very fuzzy at this point. There are some common points in the published visions about the 4G networks. In particular, there seems to be a common agreement about 4G being less about new air interface development (NTT DoCoMo 4G vision) and more about the interconnection of the existing wireless infrastructures (the vision of the European Commission). In the same way as the Internet has interconnected the deployed local networks, the 4G is supposed to interconnect the deployed radio access networks, broadcast media and perhaps wired world. However, this predicted interconnection is much more than simple network-to-network connectivity. A global harmonized mobile services network has to be designed where services are offered independently of the actual physical access method. A compatible service playground means here that if a particular physical network can not provide a service, we should be able to provide a variant of this service (of lower quality, etc.) or at least a notification (error message, explanation, etc.). Both presume a consistent user-to-network and network-to-network signaling.

Some 4G visions go beyond the essential goal described above, explicitly putting emphasis on machine-to-machine communications, ad-hoc networking or core network architectures. Different approaches have been proposed so far in the research. In following we present the main directions of these propositions.

IV.3.1. Evolution vs. Revolution

One of the mostly discussed questions regarding the shape of the 4G remains its nature: will it be a completely new, revolutionary approach or will it be a natural evolution from 3G? This topic is often discussed independently for the user-to-network interface (user link) and the core network architecture.

Probably one of the first published 4G visions appeared in [49]. Using the 10 years argumentation (described in IV.1.2) and giving an overview of the existing wireless networks, the author takes into account the steps made from 1G to 2G and then to 3G, in order to understand the next probable step. He defines the main criteria of 4G as an architecture integrating all systems, offering all services, all the time. 4G has to be prepared to support multiple classes of mobile terminals. The author presents three very different visions of 4G: the vision of NTT DoCoMo basically concentrating on 4G as a prolongation of 3G with higher data rates, the vision of the European Commission that accentuates the need for access provision over different network types (public/private, wide-are/indoor) and provisions for the growing role of M2M communications, and finally on the German VDE 4G vision explicitly willing to integrate private and public networks (namely WLANs) and extending 4G to the ad-hoc area. The author's main statement is that 4G has to focus on users and not continue to focus on operators or providers. According to the author, a 4G user is not likely to be "owned" by a provider but rather looks for seamless uncomplicated service provision everywhere.

A definition of 4G as networks of networks is also given e.g. in [51]. In their umbrella document, the authors outline the main 4G idea as a vision based on five elements: fully converged services, ubiquitous mobile access, diverse user devices, autonomous networks

and software dependency. The authors discuss these goals from four different angles: the radio aspect referring to available frequencies, regulation, etc., the user aspect referring to configurability, the network aspect concerning adaptive signaling and the software aspect regarding middleware. As main characteristics of 4G networks the authors name the predominance of machine-to-machine communications, location dependent services, privacy and security, IP-QoS and better air-interfaces (improved coverage, spectrum usage, bandwidths). A personal mobile assistant (PMA) is likely to be the "killer-application" in 4G.

In [52] the author discusses what might be real and what is hype in the initial 4G discussions. The author presents some major challenges like physical restrictions and different possible approaches (evolution vs. revolution, standardization, etc.). The author points out that unlike in 2G or 3G, which mainly aimed to air-interfaces, the core architecture has to be specified in 4G.

In [54] the authors try to identify the key research issues for 4G infrastructures and define three major scenarios describing possible telecom futures. Using this scenario-based approach they identify suitable future research topics, namely broadband OFDM air interface design, smart antennae, wireless infrastructure architecture, wireless resource management in multiple-operator infrastructures, seamless IP mobility support for mobile applications and other research challenges like asymmetric wireless infrastructures, one-stop shopping and infrastructure deployment strategies.

Thus the question of whether the path towards 4G will be revolutionary or evolutionary is not completely clarified. However, for us the development of the user link seems a natural highly challenging engineering process that has however so far proven to be reliable in term of its technological development. Thus, it seems quite normal that the development in this direction will follow the natural research and development path. The paradigm change is likely to occur in the core network and services infrastructure and in the service access. This is where the main challenges of the 4G lie. Heterogeneous architecture design, user service provision over multiple interfaces and multiple networks, and the transition to a user-centric architecture with a common high-speed data transmission core should be carried out by prudently integrating the different networks in a common umbrella environment.

IV.3.2. Possible Approaches to 4G

A description of three theoretically possible 4G schemes can be found in [53]. On a very high abstraction level, the authors describe a possibility of co-operation of heterogeneous networks. This can be principally done by one of the three following approaches:

1) Multimode devices.

Multimode devices (which already exist on the market, e.g. GSM/DECT phones, PDAs with 802.11 WLAN, Bluetooth and GSM access modules, smartphones with Bluetooth capabilities, etc.) easily expand the effective coverage area managing the co-operation issues by the installed software. This concept does not require any additional complexity in the wireless networks. However, the terminal equipment has to integrate operational logics including not only every technology-specific treatment but also the translation of quite different technological parameters to be able to make decisions. It is not clear if this can be done for multiple, very different technologies taking into account the vertical (in the sense of the ISO/OSI model) complexity of QoS, security and mobility management.

2) Overlay networks.

Another possibility is an installation of the overlay network of access points supporting all available wireless networks. The complexity here lies in the overlay network. This has to define the necessary signaling and to translate it to the corresponding functions of the underlying network technologies. Besides the physical access to the used technology, the wireless device has to implement the overlay access module that will define and use the whole 4G signaling, 4G management, 4G security, etc.

3) Common access protocol.

The third possibility is to unify the access protocols of the wireless networks, thus enabling the user to access their network by some standard means. This possibility means that it is necessary to separate the data and the control planes. Further, it is necessary to identify technology-specific functions which are part of the control plane. These functions have to be externalized and reflected to an abstraction layer/abstraction API which could then implement this common access protocol.

These possibilities are complete (meaning that there are no other different approaches to an integrated 4G system in the sense of the previous section). However, they are not necessarily mutually exclusive. It is imaginable to have some combinations of these general high-level approaches in a final solution. In the following we present some of the proposed 4G architectures classifying these according to the scheme above.

IV.3.3. Related Work on 4G Architectures

In [52] the author discusses the currently most popular approach to 4G. This approach is based on a common Internet core for different networks, unifying everything over IP and the related IETF technologies. With respect to this so-called All-IP (sometimes Full-IP) approach, the author briefly discusses the possibilities and the deficiencies in the concerned IETF protocols including Authentication, Authorization, Accounting framework (AAA), Mobile IP, IPv6, IPSec and SIP. The author points out that this approach is straightforward but also problematic in terms of QoS, security and mobility management.

The presented All-IP idea is the current state of the art approach in high-level 4G research. In the classification given in the previous section, All-IP represents an overlay network approach. IP network is used as the overlay that integrates different technologies. IP technologies are used for both control and data planes. IP base stations are used as access points in such 4G vision.

In [59] the authors research a possible core network design for 4G systems. Describing the current situation of the telecommunications and the predominance of IP-based applications, they give an outlook on estimated traffic in the future generation of wireless systems. Then they discuss possible wireless transmission characteristics in terms of transmission bit rate, spectrum, area coverage, hierarchical service area and define the network requirements as seamless connections, reduction in the number of control messages, short delay at handover, reduction of cost per bit, service integration based on IP and movable network support. The network architecture is then defined as a core network (CN) connecting different access networks like 4G-RAN, WLAN, 3G and PSTN to the Internet. CN and 4G-RAN are completely IP-based. The terminals have IP-addresses assigned. The CN is directly connected to 4G-RANs and the Internet and uses gateways to connect to the Public Switched Telephone Network (PSTN) and 3G. Mobility management is done by using the hierarchical Mobile IPv6 approach.

Additionally, the article discusses some issues in the 4G-RAN configuration. Thus, this proposal is an instantiation of the All-IP approach.

Another All-IP proposal is discussed in [60]. The recognized requirements here are huge (IP-) multimedia traffic handling, advanced mobility management (MM), diversified radio access support, seamless service and application service support. The authors then discuss possible solutions for MM and seamless services and name Mobile IP, Cellular IP and similar techniques. However, they recognize the deficiencies of such systems since they are hardly suited to provide a mobility management of the same quality as is the case in 3G. The authors claim that the networks beyond IMT2000[6] should be much more location-registration-oriented and should identify the location registration management as a study topic. For instance, hierarchical or concatenated location registration techniques have to be studied. Then they discuss handover issues distinguishing local handovers and overall network handovers and identify this feature as a further study object.

Trying to provide an infrastructure-independent access to services and applications for highly mobile users, the authors of [61] present a communication gateway based solution. Originally driven by an automobile environment, the basic idea is to install an intermediate element between the actual user equipment and the serving networks. Such a communication gateway thus resides within the end-system. Including caching and switching units, the gateway provides a general middleware interface to the applications. Thus, this approach pushes the intelligence towards the end-systems, trying to map user requests at their origin to available networks and services. In our classification, this proposal thus represents the multi-mode device approach.

The authors of [56] take a slightly different approach. Mainly dealing with QoS support over different wireless infrastructures, they define a new intermediate layer between the IP and the 2nd layers. This Wireless Application Layer (WAL) then provides a QoS-generic interface for IP featuring uniform guaranteed link reliability and traffic control. The position of WAL in the ISO/OSI layer implies a hop-by-hop QoS agreement logic. The details on the modular architecture of WAL, its class and association based QoS provision, Snoop TCP method to avoid congestions in the TCP layer can be found in the paper. In our classification this proposal is still an overlay network proposal. The overlay is built by WAL instances that have to be integrated in the terminals and in the access points. IP is then used in the All-IP manner, but the heterogeneity is hidden within the WAL that acts as a convergence sub-layer but. WAL instances rely on SNMP to collect the necessary decisions bases etc.

The user verification and network access in heterogeneous environments represents one of the major 4G problems. This is discussed later in details. One of the problems is the access protocol but there are only some open questions concerning the backend trust architectures and multi-domain, multi-party authentication, authorization and accounting.

An interesting related work seems to be [62]. Introducing the concept of a so-called *virtual operator*, the authors describe how an authentication service reachable over the Internet could authenticate its users in a foreign hot spot environment using AAA. As potential virtual operators the authors see Internet Service Providers, Content Providers, cellular operators or pre-paid card issuers. To reduce the number of necessary trust relationships between potentially numerous hot spot operators and diverse virtual operators, the authors propose a commonly trusted broker entity.

IETF currently works on the Protocol for carrying Authentication for Network Access [63] in its PANA working group. PANA specifies an architecture very similar to the

[6] International Mobile Telecommunications 2000, ITU's common name for different 3G variants.

IEEE 802.1X architecture used in this work for LAN/WLAN access. PANA is link layer agnostic transporting authentication information between the PANA client and PANA authentication agent at higher layers. Since it is principally capable of identifying users, PANA could thus be used as a common access protocol to heterogeneous networks. However, since PANA has to access a higher level element, the L2 mostly remains unprotected. Also, after the (unprotected) L2 establishment, the local PANA client needs to discover its network's pendant, the PANA authentication agent (PAA). This involves discovery broadcasts and round trips. PANA here nicely illustrates the problems inherent to higher layer network access: questionable security, holes in the access controllers, broadcasting in the access phase and high network access latency.

Besides, PANA does not optimally support mobility: without additional mechanisms, the authentication has to be completely restarted at the next visited PAA (even within the same network). Such mechanisms could be a L3 (i.e. in the 4G scope typically IP) context transfer protocol that would allow arbitrary context transfers between different PAAs. IETF will shortly publish its CTP (context transfer protocol) specification [64] as an experimental standard. However, the payload formats for CTP have to be specified too.

The work on the public access wireless networks (PAWNs) can be interesting in the 4G scope since it has to practically resolve several problems very similar to the anticipated 4G problems. PAWNs are typically implemented with IEEE 802.11 technology. Since the integrated 802.11 mechanisms are insufficient for almost all typical PAWN areas (per user quality of service, system-wide mobility, security, user network access, etc.), the solutions proposed for PAWNs are typically completely decoupled from the underlying technology. Hence, the practical experiences gained in such installations are of tremendous importance for the 4G research.

An approach for WLAN hot spots providing a secure wireless Internet access in public places is Microsoft's CHOICE [65]. The authors build a network that globally authenticates users and then securely connects them to the Internet via a serving 802.11 WLAN. A reasonable argumentation against IPSec for this purpose can be found in the publication. Introducing a new software module (PANS) instead of IPSec, the architecture promises authorization, access control, privacy, security, last hop quality of services and accounting. However, this software (responsible for packet marking on mobile hosts) has to be installed on all mobile terminals, effectively modifying protocol stacks. The WLAN itself is open but does not allow any connections to any other networks, except for HTTPS connections to the global authenticator (global MS Passport service) and HTTP to the local web server where e.g. the software module can be downloaded. Network's PANS authorizer module obtains key information from the global authenticator after successful user authentication. The authorizer can also install all required policies. It then reroutes the traffic to a PANS verifier. The latter actively processes every packet checking the mark/tag added by the PANS module running on the mobile and providing e.g. per user access control and accounting.

Mobility support for public wireless LANs is presented in [66]. Using a similar packet tagging approach as in CHOICE, the authors describe their GUIDE/GUIDE II systems. Originally meant for a metropolitan scale access using modified client protocol stacks, GUIDE offers ordinary citizens secure and accountable Internet access over the deployed 802.11 WLAN-infrastructure. GUIDE II adds handover management using Mobile IPv6. IPv6 datagrams are tagged by clients using the modified MobileIPv6 stack. Programmable access routers ensure that only packets containing valid access tokens get to the trusted core network. Over an access router, users authenticate at an AAA authentication server. The latter distributes session keys to the access router group and the

mobile terminal. User payload encryption is optionally possible between the router and the user equipment.

In the following sections, we try to summarize the most typical problems in the 4G scope recognized so far within related ongoing projects and related work.

IV.4. Challenges Implied by 4G

In the heterogeneous 4G world, different systems have to collaborate to provide users with the expected services. Such collaboration is only possible if some level of understanding is guaranteed. This can be achieved by standardization, rules, common APIs, etc. or by some system part complex enough to adapt to all needs. In 4G, a reasonable compromise between the homogenization and the feasibility on one hand, and the solution complexity and the cost on the other hand has to be found in order to achieve an adequate system performance.

IV.4.1. Security

Current wireless technologies have different security considerations and provide corresponding security definitions in the standards. The latter are naturally dedicated to the device security and thus define protective measures on the respective link layer. In 4G, different link layer technologies are likely to be combined. Also, the focus changes: in the personal communications the security focus should be on users, not on devices.

The problem with the 4G security is twofold. On one hand, there are very basic open questions that have to be answered by the ongoing research by weighing practical constraints against the required security level. But what is security in 4G, if we do not know how 4G looks like? The system architecture plays an essential role for every security definition. For instance, if users are not “owned” by providers [49], how can trust be established and to whom? A consistent security policy has to be defined along with the architecture identifying subjects, objects, trust relationships, authorizations, risks and protective measures. This is however difficult (keyword: *heterogeneous security*). On the other hand, there are practical problems concerning the compatibility of solutions. In the best case, the existing security systems are limited to the identified needs. In the worst case, these fail to provide the necessary protection, typically because of conceptual or implementational flaws. How can the defined security policy for the entire system be applied to the deployed entities given that the available solutions are different and limited to a part of the system? For instance, if the security policy identifies link encryption as a necessary confidentiality implementation, how can this be activated and with which keys and properties? How can it be guaranteed that the strengths of the different encryptions in different technologies are adequate? What to do with the technologies that do not provide link encryption? The security policy must consider these cases and provide answers to such questions.

The aforementioned practical problems can be avoided if the technology-dependent security measures are not used. Instead, all security measures could be applied in the overlaid technology. However, it is often not possible or at least inefficient. For example, 2G/3G network providers rely on L2 security measures for network access control and link encryption. While the link encryption is not important for the provider, the access control is primordial for infrastructure protection and revenue guarantees. Moreover, the L2 security measures are often implemented in the network interface hardware. Their design includes power consumption and computational resource considerations. A higher

level solution would be implemented in the device control logics, i.e. typically software. Given the constraints with the 4G terminals (wireless security processing gap, see Section IV.2), it would be wise to use the hardwired security solutions in the network adapter. Furthermore, in the strict OSI logic, multiple links could lie between the user and the used L3 device (router), but only one link is possible between the user and any used L2 device. Thus, the L2 security measures are guaranteed to be implemented in the first network entity (the access device), i.e. next to user, at the very edge of the network. That brings the security as close to the user as possible and thus guarantees physical infrastructure protection. Moreover, it principally scales better since the access devices are designed to support a fixed number of connections. Another point is that higher level security solutions can not achieve the same user privacy. For instance, user location privacy is in danger since lower layer addresses (such as world-wide unique MAC addresses) can not be hidden by higher layer security measures[7].

For the reasons stated above, we think that L2 security is indispensable in 4G.

In this scope, a particular security problem is bound to the user network access. The 4G user has a terminal with multiple network interfaces. The security measures for each interface have been designed according to an initial security analysis during the technology standardization phase. Since the technologies are meant for different purposes, the risks and the defined security functions are likely to be different. The security mechanisms are definitely different. Thus, every interface has different requirements on credentials in terms of identities, credentials, etc. These requirements have to be fulfilled since otherwise the interface could be unusable or the access by the means of this interface impossible. If the user definition in the system is consistent, then the 4G user can not be expected to use multiple identities: in 4G, every network provider needs to be able to identify any given user correctly, in particular in the different access networks which the user might be using simultaneously. That is important for the authorizations defined in the security policy. It is equally an important requirement for correct billing. Network access can thus be divided into various sub-problems which are treated in more details in following.

Network Selection

Users must be able to collect information on the access networks of all available providers. Most importantly, this is required for the decision of which network the user should connect to. For instance, it can not be generally assumed that every network is accessible for every user (e.g. because user's home provider does not have any roaming agreement with the provider of the detected network). *Network selection* is a problem in 4G since the network identifiers do not necessarily exist. They also have very different meaning in different technologies. If a 2G provider wants to deploy a supplementary data service over a 802.11 WLANs, what should be used as a network identifier? There is no regulation on SSID naming in the 802.11 WLANs. Besides, in the various 4G technologies with the very different proposed services it is difficult to believe that a network identifier alone is a sufficient base for the network selection decision.

User-Network Authentication

After the information collection, some networks can be eliminated by policy or user wish (e.g. a pre-configuration of the type "never use provider X" or "always choose the cheapest available service", etc.) Now, the user can access some of the available

[7] Although the lower layer address and the user identity are two completely different identifiers, one initial passive network observation in the proximity of a victim allows an establishment of a direct relationship.

networks. The L2 user-network authentication is a problem in 4G since the logical and technological requirements are very different from technology to technology. We illustrate this on an arbitrary example, comparing UMTS and standard 802.11 security.

UMTS uses an external module (USIM) that hides the actual authentication method from the used device and the visited network. The authenticated logical entities are the USIM and the visited network. USIM is supposed to grant access to the network to the device (i.e. also to the user). The USIM is capable of key derivation after a successful authentication.

Basic 802.11 defines a handshake procedure based on a shared key. The whole procedure (i.e. the authentication method, the exchanges, the cryptographic functions and the success conditions) is hardwired in both network interfaces. The only authenticated entity is the network interface of the user device (i.e. the access point is not authenticated). The authentication does not derive any key material. Moreover, this procedure is completely useless because of a concept error (see Appendix A).

As can be seen, the provided services are very different in terms of capabilities and the achieved security level. Even if today other security models and methods are available for WLANs, this underlines the problems in the general 4G scope. The resolution of this problem must not lead to security problems. Thus, if the L2 authentication is to be used in the 4G scope, then it has to fulfill some minimum requirements. Otherwise, higher level security has to be used and the associated higher level access controllers have to be collocated with the L2 access devices. If that cannot be guaranteed, this technology should be considered unsuitable for 4G.

From today's perspective, the requirements on the L2 authentication are the cryptographic strength, mutuality and dynamic key material negotiation for the subsequent session protection. The key material negotiation should provide *perfect forward secrecy* (PFS), i.e. a successful attack on the produced key material should not give any clues on the long term secret such as the used credentials. User location privacy should be supported, i.e. if possible, any user-specific identifiers should be unreadable for a third party.

Note that we do not formulate any requirements on the authentication logic (how many parties involved and how), used protocols, implementation, method placement or on the used trust representation.

Link Encryption and Data Integrity Functions

Different wireless systems use very different link encryption and data integrity techniques based on different mechanisms. Typically, shared key mechanisms are used for both link encryption and data integrity. Very often proprietary solutions are implemented in both cases. The needs of the used encryption and integrity functions in terms of key properties (format, length, known weak keys, etc.) and the optionally used initialization vectors are very different. The provided security levels are also quite different. Thus, the situation of these functions is similar to the user-network authentication. If some minimum requirements cannot be fulfilled, these could not be used.

Simultaneously, both functions are in use during the whole session. Thus, the power and resource considerations of these are especially critical. For that reason, we think that both encryption and integrity functions should be implemented in the associated network adapter (hardwired or in form of hardwired cryptographic bricks connected by the softwired firmware definitions). Both functions must use the key material derived during the authentication session and support rapid re-keying, both periodical and on-demand. Ideally, both functions should be cryptographically strong. However, if flaws are

detected, the rapid re-keying can help mitigate the problem by changing the encryption keys very often.

IV.4.2. Mobility

As has been discussed in III.4, mobility management is a technological challenge increasing the complexity of modern wireless systems.

Mobility management in heterogeneous systems is much more complex since different technologies have very different provisions for mobility support. On one hand, the system mobility management has to cater for the deficiencies of some technologies that do not integrate the necessary minimal mobility support. On the other hand, the mobility management has to provide means for the inter-technology mobility with the same or comparable parameters (quality, security, etc.) as the integrated solutions. In a 4G system, we thus have to technologically distinguish an inter-technology handover, a *vertical handover*, from an inner-technology handover, a *horizontal handover*. The goal is to provide a free network choice (roaming access) and service session continuity (seamless service provision). Principally, this could be implemented by device mobility or session mobility support. In some cases, it could additionally require network mobility support.

Seamless mobility within the existing cellular mobile networks (2G-3G) is already implemented and actively used. However in Europe, mobility and roaming between different coexisting cellular networks is usually prohibited by the national telecommunications regulation to encourage the development of a gapless national coverage.

In the case of WLANs, the situation is rather different. Seamless mobility within a WLAN is typically possible within the same L2 subnets (see IV.1.1). However, WLANs do not natively define any inter-provider roaming provisions. National WLAN coverage is not realistic and not necessary since the 3G technology is technologically superior in broad deployment (more spectrum efficient, integrated management, etc). Thus, user roaming between the different providers' WLANs is likely to be admitted by the regulation authorities. Since WLANs are exempt from any other form of regulation (spectrum licensing, etc), everybody can deploy a WLAN. We thus presume that different overlapping WLANs can co-exist in outdoor business areas and, in fact, they already do (e.g. in Paris at the Luxembourg metro station a café hotspot overlaps with a hotspot of a local university). The user thus faces new choice opportunities. In this manner, local providers will nevertheless be in a competitive situation. A reasonable amount of cooperation and fusion will probably lead to providers extending their activities to at least regional levels. Such specialized enterprises can use their experience to entertain and manage numerous hotspots. The evaluation points for such a provider will be the provided services (e.g. Internet access, VoIP, etc.), their respective quality (e.g. throughput, latency, jitter), prices and, finally, the coverage area, i.e. the number of sites where the user can get this service. Hence, on one hand providers will be interested in local site management, resource control and resource usage adjustment. On the other hand, providers will try to extend their serviced area (more customers, more income) and will thus cooperate with other providers, thus attracting users with an advertised bigger coverage area.

Mobility management in an interconnected UMTS/WLAN infrastructure is discussed in detail in [67]. Using AAA and MobileIPv6 and considering possible QoS management (i.e. no techniques are used but IntServ, DiffServ and MPLS approaches are considered), the authors take a detailed look on the necessary signaling. They propose to treat the micro-mobility using the respective network technology. The macro-mobility using

MobileIP between UMTS and WLAN and vs. features seamless handovers by using an intelligent mobile terminal MobileIP agent. Authentication and authorization, which are crucial in that approach, are left to AAA. Using these ideas, the authors additionally aim to provide mobile-initiated location based services using a Java interface. Once connected, a mobile node can use its generic location service interface and ask the network about its own (and only its own) position which will be returned by a central location data-base in GPS format. Clearly, this information can be used by applications.

An interesting comparative work on existing UMTS/WLAN interconnection strategies is presented in [68]. The authors discuss three basic ideas of how UMTS and WLAN could be reasonably interconnected in order to provide seamless roaming. The first idea follows the All-IP approach: based on IP, Mobile IP components have to be installed on the mobile nodes and in both networks (foreign agents/home agents). The authors state that the handovers suffer long delays and experience high packet loss. Additionally, Mobile IP's typical triangular routing [31] is clearly suboptimal. The second approach is a gateway approach, introducing a new logical instance interconnecting the networks and responsible for signaling translation. That helps to separate the operations of both networks since both networks can be operated independently (this is sometimes referred to as *loose coupling*). The handovers are faster and fewer packets are lost compared to the Mobile IP. However, the implementation of this approach is not straightforward: some UMTS protocols should be refined and AAA/HLR information exchange precisely defined. The third approach is an emulator approach, where the WLAN acts as another radio link to the UMTS base station. This approach completely reuses the UMTS signaling and provides UMTS-like mobility without Mobile IP modifications and with shortest handover delays. However, it tightly couples both networks and the Internet traffic from mobile nodes has to be routed over the gateway GPRS support node (GGSN). Standardization is necessary for the realization of this approach (often referred to as *tight coupling*). Studies are currently pursued by ETSI for HIPERLAN and UMTS (ETSI BRAN Project). Furthermore, the authors present comparative simulation results for handover latency that confirm their analysis.

Soft handover remains a hard problem to resolve in the All-IP approaches presented above. The problem is that multiple streams of the same IP traffic have to be distributed via multiple base stations to a mobile station. Then, the arriving pieces have to be identified as copies at the mobile station in order to be combined in the right way although they arrive from different sources. Obviously, a similar problem has to be resolved in the inverse direction, i.e. from the mobile to multiple base stations. An interesting publication in [69] identifies these problems and proposes solutions designing an IP-based base station (IBS) and defining a new IP procedure for content synchronization. The authors correctly point out that the existing IP-mobility techniques do not provide any soft handover solutions. Moreover, they claim to have found nothing on soft IP-handover in the current literature. Discussing the problem of MN-to-IP-subnet assignment, they note that if MN can keep its IP-address while changing IBSs then all IBS have to belong to the same IP subnet and therefore every IBS will assume that it can reach the MN. On the contrary, if every IBS manages its own IP-subnet, additional IP-address change delays occur during the handover increasing the overall delay. The authors propose a proxy-ARP based approach (which they call "shadow address approach") to enable one IP subnet for multiple cells without otherwise necessary multi- or broadcast. By using matchable L2 streams coming from different IBS they want the radio software part of the MN to automatically recognize the frames as duplicates and thus provide data content synchronization.

A different approach is described in [70]. Without trying to provide soft handovers, the authors plan to provide a paging service and fast handoffs in the All-IP networks of 4G.

The authors present their own mobility protocol, IDMP, and compare it with Mobile IP and Cellular IP approaches. IDMP introduces a hierarchical mobility approach assigning a local and a global care of address (CoA) to each mobile. The global CoA represents the position of the mobile seen from outside of the domain, during the local CoA changes with the intradomain handovers. Thus, latency-prone binding updates to exterior correspondent nodes are not necessary in micromobility cases. The authors describe how the mobility agent in their approach can enable a fast handoff mechanism for a local handover within 4G. The mobility agent (i.e. the host intercepting all messages addressed to a mobile node) gets the message about imminent location change either from the mobile or from the base station. On reception of this message it starts multicasting all packets for the mobile to the set of neighboring BSs. These BSs buffer the incoming packets and forward them to the mobile if it enters their cell. Thus, no packet loss occurs. The delay is minimized since the packets already await the mobile at the base station after the reattachment. Additionally the authors present a paging service (which is missing e.g. in Mobile IP). Paging can help minimize the number of necessary binding updates or address renewals since the mobile can detect the need for such an update by receiving the paging message within the new cell.

It is currently not clear how the global mobility management should be managed in the 4G architectures. Since Mobile IP based solutions seem to have performance drawbacks, newer paradigms propose to achieve the overall goal by combining L2 and application level mobility support. Local mobility is left to the integrated access technology solutions. These can provide efficient handovers since no provisions for higher layer reconfiguration are required. Global mobility is managed on the application layer, at the user level. This can be done e.g. by the SIP protocol [32]. Hence, this paradigm represents a compromise between session and device mobility (L2). However, with this approach, global mobility results in quite slow handovers and the application has to provide mobility support, trying to softly change from one technology to another. Unfortunately, with this approach, the changes of the access technology within the same provider network also trigger the same mechanisms as for the really global mobility.

IV.4.3. Heterogeneity of Signaling

As has already been stated in III.6.2, mobility and QoS in wireless networks are challenging issues. Both can be solved in homogeneous systems (as proven by 2G); however this is done by an integrated management using a signaling system of high complexity (SS7).

However, in 4G the used subsystems are heterogeneous. For instance, in 4G an end-to-end QoS (parameter negotiation and channel reservations) has to be managed by the two mobile ends over potentially different access links, i.e. featuring completely different quality properties (bandwidth, delay, jitter, BER, etc.). Given that the QoS establishment on every connecting channel is an issue per se, 4G here introduces a new complexity dimension. To be able to make decisions on the sufficiency of the quality of the end to end connection, the mobiles need to normalize the measured link conditions by some metrics. This normalization has to be consequently translated back into appropriate parameters used in every part of the whole system, assuming the participation of all provider networks and the core network. However, different technologies provide very different characteristics and these are difficult to compare.

Another problem is the mechanism for such negotiations. Both management and link signaling are heterogeneous. In other words, a common signaling system does not exist. Some architectures hardly define any user-related signaling (i.e. only basic access signaling is defined). Others have a complicated signaling subsystem which is available

to the network adapter implementation. This has to be translated in order to be used in the interface-spanning control process.

In the All-IP approach, IP could transport the signaling exchanges. However, this represents a burden of additional delays necessary for access router discovery, IP-layer configuration, etc.

Mobility, and especially seamless mobility, considerably worsens the situation introducing handover delays and requiring context transfers between technologically different networks. Note that context transfers are probably unavoidable for seamless mobility due to relatively high end-to-end channel capacities. However, they also mean that mobility can not be done by terminals only, since they enforce either network device participation (common signaling?) or introduce dedicated mobility servers in every participating network (complexity, scalability, etc.).

IV.4.4. Heterogeneity of the Available Services

The basic service access in the 4G is to be based on IP since user data are expected to be transported over IP exclusively. Thus, the completion of the IP access is a necessary condition for service access. IP configuration can only be started after the successful L2 access. This adds to additional network access delays and can become critical for the seamless mobility support.

Moreover, in contrast to the homogeneous 2G/3G networks, the service environment homogeneity can not be assumed. Instead, we expect every operator to implement very different service sets in the respective infrastructure. The access permission to the services over any given access technology depends on access technology suitability for the service (e.g. voice calls over 802.11 cannot currently be supported with a satisfactory quality), the user identity (authorization) and the current network situation (available resources). Since the user could access different provider networks over different technologies at the same time, a robust and rapid *service discovery* is necessary on the user link.

Service discovery is very challenging because of the heterogeneity and the dynamicity of related networks. Existing service discovery protocols for wireless networks are usually classified in two categories: *proactive* (or push-based) and *reactive* (or pull-based) SDPs. A proactive SDP uses a broadcast mechanism, where each node advertises the services it can offer to the rest of the nodes in the network. This push-based strategy is used e.g. in IEEE802.11 and HIPERLAN. In case of reactive SDPs, each node queries other nodes for the services it requires. This approach is used e.g. in Bluetooth networks. A problem with proactive SDPs is the additional steady bandwidth requirement for message broadcasts. Besides, there is a potential for additional latency in discovering services because nodes send the advertisements periodically. In turn, reactive SDPs suffer a constant latency to find the required services. This latency corresponds to the time the broadcast message needs to reach the offering node, the processing delay and the response time. In that sense, proactive SDPs offer more design flexibility since they represent a tradeoff between the overhead and the performance. A proactive SDP can be faster than a reactive SDP, however at the cost of a high bandwidth overhead. This can be an important issue for wireless links (since the resources are considered scarce).

Service discovery protocols can be based on IP; however, the usage of a higher layer protocol like IP typically results in additional higher layer configuration latencies as already discussed in Sections IV.4.2 (mobility) and IV.4.3 (signaling).

Furthermore, the availability of certain services is one of the most important criteria for the network selection decision. Thus, we introduce the term *early service discovery*. Early service discovery means that users should be able to discover services and their respective characteristics before the actual access to the network, i.e. to be able to make the decision. That is impossible if the network access control is carried out by L2 but the discovery is based on L3. Early service discovery is thus a heterogeneity problem in secured infrastructures.

IV.4.5. Payment and Accounting

It seems obvious that provisions for a coherent and precise billing are indispensable in the 4G context. However, firstly suitable business models have to be evaluated. In the voice networks users pay per call. But what should users pay for in the multi-services network? Today, the usage of Internet for distant calls results in major cost saving. However, with the ongoing network convergence and the open competition, this cost saving has to vanish. In a open market 4G system, standard economic models apply. Thus, each call will have some price justified solely by the demand and offerings. Accordingly, the price will depend on the provided quality but not on the used technology.

The requirements derived from these business models have to be applicable to the provider networks and to every access network. This however represents a considerable technological problem because once again the technology-dependent signaling has to be treated in an inter-technology manner. The overall provided service used over different networks basically represents an advanced service from a charging perspective since different contracts (client-network, network-network) are used for every service access. The price could be variable in time. For some general remarks on billing refer to [71].

Session accounting information has to be gathered in the network and service access controllers for billing purposes. Exact accounting start and stop conditions should be defined in every technology. User devices should log the accounting data independently to have a verification base. For 4G network access, the L2 accounting can be a problem if it is performed with user participation (e.g. for non-repudiation). In that case, signaling is used on L2 which results in the heterogeneity problems.

In this work, we do not treat this complicated problematic in detail. However, where applicable we try to respect the possible requirements. The proposed mechanisms principally allow for accounting and billing. An interesting work optimizing billing mechanisms in multiparty scenarios in future telecommunications is [72].

IV.5. Conclusion

In light of the popularity and the proliferation of the Internet technology in the core telecommunications businesses, there seems to be broad support for the convergence towards a unified IP-based architecture. Given the progress in the router development and the ongoing bandwidth increase in the modern Internet, we believe that All-IP is a valuable proposal. In particular, it seems that IP is well suited for the bit-transport fabrics built by the service and the connecting core network.

However, as stated in [52], there are serious considerations with regards to diverse management tasks in All-IP architectures; mobility management, security, and QoS mentioned above. Although the academic research in these domains has progressed a lot, most tasks in today's practical networks are carried out by other means, i.e. on the layers below or above IP. For instance, network providers manage their infrastructures by

MPLS [73] allowing the creation of customized IP packet delivery paths. 2G and 3G operators use the integrated 2G/3G management possibilities to provide IP connectivity, but IP as such remains a user service. In GPRS and UMTS core networks IP is used in the data plane only. At the moment, it seems unlikely that major operators will use IP-based management, for diverse reasons. IP Mobility management [31] has difficulties gaining momentum. In the same way, the diverse IP QoS management proposals like IntServ [74] or DiffServ [75] mostly remain an academic issue. IP Security (IPSec) [76] principally does not allow for physical infrastructure protection, unless all access points implement IPSec for access control. Besides, from user's perspective, IPSec is too low in the OSI stack – it is on the device level securing logical appliances and not users (i.e. sessions). In modern networks, security is typically implemented by providers at L2, to assure physical infrastructure protection and network access control, and by users at L5/L7 (e.g. TLS [77]) to enable secure service access, data confidentiality, etc.

Hence, the access technologies have to fulfill certain requirements in terms of integrated L2 mechanisms. This is one of the conclusions of the security discussion. The same conclusion has been underlined in the discussion on the available services and their quality. However, there are two major issues with the low layer definitions. How to provide a unified system access for users over different technologies, and which technologies fulfill the requirements on 4G access technologies?

Being maintained by the national scope telecommunications companies, 2.5G and 3G networks provide a roaming service, different quality of service levels and well-defined billing mechanisms. The authentication is based on a so called subscriber's identity module (SIM in GSM [78], USIM in UMTS [79]) which stores the credentials and the authentication algorithms. So, this type of networks is relatively well prepared to serve as an authoritative domain in a multi-provider context. In the All-IP approaches, provisions have to be made for IP stack installations on the terminals. As mentioned, national roaming is to avoid due to the current regulation.

In 802.11 WLANs, the work on all three points is in progress in different working groups (WG) at the IEEE. The concerned WGs are:

- 802.1X WG defines secure port access control mechanisms for 802 networks with implicit provisions for roaming [80],
- 802.11i WG proposes a security enhancement for 802.11 WLANs [81],
- 802.11f WG proposes a secure context transfer protocol between the access points allowing for smoother mobility by supporting faster local handovers [82],
- 802.11e WG currently works on different QoS aspects on the 802.11 links [83].

The results of these working groups are partly already available in form of standards, drafts, etc. and should be used in the 4G scope to fulfill the requirements on 4G links.

IP becomes the technology of choice as a unified data packet transport mechanism. However, we state that today almost all control plane tasks in the private provider networks are carried out by the mechanisms available in the used infrastructures, independently of IP. Unfortunately, if the used mechanisms are too device-dependent, these can not be used in 4G with its heterogeneous access networks. If the mechanisms are too standardized, providers lose flexibility as they are limited to the same solution. It seems crucial to provide a flexible management architecture to support as many different provider types as possible. Simultaneously, the providers should be able to cooperate to enable system-wide user roaming.

A global solution has to be constructed out of link solutions. The collaborating providers have to install sub-systems in their networks enabling support for the agreed features. The

exact architecture of such systems has to be defined with management and security in mind. The challenge is to design a system architecture that is practically feasible and sufficient in terms of security, mobility support and QoS.

In the next chapters we address several issues within the All-IP architecture. More precisely, we address user roaming access issues, describing the provider-provider and the user-provider interfaces. In the subsequent chapter, we design different flexible management infrastructures for the private provider networks.

CHAPTER V
4G Logical Network Access

In this chapter, we define our global 4G network architecture and the necessary entities. This architecture and the nomenclature are used consistently in the rest of this document.

Using this nomenclature we define a common access model to networks and services. We then discuss trust and mobility in our model. In particular, we study user roaming support capabilities of IEEE 802.11, discussing trust relationships and system architectures enabling transparent access to different networks. We propose optimizations to the standard user roaming, achieving a faster and more efficient inter-domain handoff.

We then study 4G network discovery and 4G heterogeneous network access. We identify general problems within the available methods that can not be easily resolved within the classical access model. Therefore, we propose and develop an extended system access model. We show how such model can be principally implemented in an 802.11 environment, discussing the advantages and shortcomings of the different alternatives. We then present our approach EAP-SIG and a working implementation. We show how EAP-SIG can be used to efficiently support micromobility in a secure 802.11 WLAN.

V.1. Our 4G Vision

Our vision is motivated by previous work (see IV.3) and ongoing development of the global telecommunications networks, in particular of the Internet. It respects the fact of the proliferation of Internet technology in all telecommunications branches and is similar to the All-IP approach when used for data transport (discussed in IV.3.3).

Learning from 2G and 3G experiences, we envisage an architecture which allows the maximum possible infrastructure reuse. The idea is to minimize a risky engagement with a particular technology and to guarantee the long term flexibility for the involved authorities. We believe that the versatility here can provide an enhanced flexibility both technologically and from the business point of view. This ultimately market-driven solution should be capable of providing any service in any manner, restricted solely by user's demand and not by any technological factors.

V.1.1. From Service-Centric to Data-Centric Approaches, from Technology-Centric to User-Centric Approaches

The classic telecommunications industry approach dominated by the national-scope telecom operators with the well-managed infrastructures can not currently provide a cost-effective focused access to the Internet services. This is particularly true for the developing countries where massive updates of the existing infrastructure can not be afforded.

In its initial collaborative work, the telecommunications industry was much influenced by the dominating demand for the voice telecommunications. The 1G and 2G systems were originally designed to provide one single service: the mobile voice telephony. Their system design was *service-oriented*. As a result, the conceived core infrastructure is circuit-oriented and the wireless link's capacity is tailored to the voice-implied bandwidth requirements. Due to these properties, 2G currently provides a reliable voice service; it is however quite difficult to reuse this infrastructure for other purposes. However, deploying a new infrastructure for every service is not scalable and financially impossible. Especially with the modern digital technologies, it is much more efficient to reuse the same infrastructure for different services.

3G development is an example of a *network-oriented design* process (sometimes also called operator-oriented design). It is a step ahead from the service-oriented design of the 2G system since it explicitly provides for infrastructure reuse for various services. Principally aiming at operators and networks, such design tries to respond to operator's management requirements. It thus produces homogeneous technologies comprising everything the operator has requested. According to this design paradigm, the 3G technologies deliver voice and data within the same infrastructure. In presence of an existing voice-oriented 2G infrastructure this renders the only added service – the mobile broadband data – quite expensive in itself. The operators have to amortize the network deployment and the license cost over the new service. Thus, from the user's point of view, this new service is often perceived as too expensive.

To be able to provide cost-effective data services at any chosen place in the world we need more user-oriented and data-centric approaches than what 2G and 3G paradigms deliver. At least in the mid-term, the hope here lies in a more opportunistic approach from the technological point of view. Indeed, the user usually does not care about who provides a particular service and how. The user cares about the availability of services, their performance (throughput, latency, etc.), quality of service (i.e. the performance and

the variation of the performance factors), the ease of use and service prices. Accordingly, the *user-oriented design* tries to respond to these user wishes assuring the possibility to freely choose an available service. Choice as the driving factor for the competition plays a crucial role in this scope since it results in better and cheaper technology.

From the system's point of view, the resulting overall architecture delivers very different services through completely heterogeneous access networks. User-oriented design has to cope with the question how to manage the system and how to provide users services with an expected quality. The management is important because a good management reduces the operational costs. The provision of the expected quality is the main factor for the user satisfaction.

Such architecture could help to achieve more infrastructural and architectural flexibility providing a free technology choice for the local operators and thus, in the final run, reducing the costs and offering more choices for the users. By featuring more flexibility, this step to further diversification gives new opportunities and could help e.g. to reduce the cost or to mitigate some aspects of the digital divide problem.

At the same time, this task is not simple technologically. As could be seen from the above examples, the service-oriented design approach is a straightforward technological way to conceive a network dedicated to the needs of one single service. Provision of more services within the same infrastructure makes it more difficult to assure that every service individually is provided in a satisfactory way. We can generally allege that the *quality of service* (QoS) in the multiservices network is more difficult to maintain because very different requirements have to be fulfilled by the same infrastructure. Yet, owing to this common homogeneous infrastructure, with the network-oriented design it is still relatively easy to conceive systems enabling a comprehensive network management. The necessary dynamic infrastructure-to-service adaptation (e.g. for QoS) can then be achieved based on the management functions.

The step to the user-oriented design potentially implies a broad diversification of data transport technologies providing different services. Thus, the resulting systems inherit the problems of the dynamic per-service QoS provision. Additionally, we run into difficulties trying to consolidate all these different technologies and make them do what the operator wants. This applies to the network management in general. In particular, it concerns the mentioned QoS provision problematic and also raises diverse security considerations, both of the operators (infrastructure control and protection, resource usage control, accounting and billing) and of users (data confidentiality, location privacy, flawless billing).

Hence, the user-oriented design opens new possibilities but potentially results in a heterogeneous environment. To be deployed and maintained by the operators, this environment needs to be understandable, manageable, flexible and secure. To be used, it needs to be user-friendly, reliable and fair. In particular, users should be able to use different services over different infrastructures in the same, familiar manner.

Thus, we need to develop more flexible infrastructures and more sophisticated mechanisms for infrastructure access incorporating but hiding the whole technological complexity. These mechanisms should provide adaptability to both users and contents. In this work we thus study the requirements of a heterogeneous network access mechanism and the necessary corresponding network management functions in the scope of the future integrated environments.

V.1.2. Multi-Provider Network Environment

For the next generation wireless networks, the accent lies on users and the requested services [49]. For the flexibility and cost reasons, the 4G architecture has to be able to integrate different technologies to provide *services* to users. Services are divers offerings, commercial or free, ranging from a basic connectivity (e.g. to the Internet) to more sophisticated services such as voice calls or Instant Messaging (IM). To provide more complex services, some providers can use services proposed by other providers.

We see 4G as a potentially open, heterogeneous, user-oriented architecture, consisting of different service and access networks. These networks are operated by different authorities. We call such authorities *service providers*[8] if access to services is possible over their respective infrastructures or *networks*. The global 4G architecture is shown in Figure V-1. It is composed of a panoply of service provider networks (SPNs) connected by an IP-based core network for any global data exchanges. SPNs principally support different wireless access networks (AN). AN technology can range from personal to wide area networks.

Each provider may, but is not required to, have own *users* and propose multiple services over different access networks. Users are defined as logical system identities subject to the service contract between two legal bodies, one representing the provider and the other representing the served user. This definition implies that every user corresponds to a *service contract* with exactly one provider[9]. Note that this contract requirement does not imply any price models or restrictions. Since every user corresponds to one legal body, we use these terms interchangeably in the rest of the document unless explicitly distinguished.

The service contract provides the trust relationship and the set of authorizations. From the user's point of view, the provider from the corresponding service contract is called *home provider*. If a provider is used only for user identification, authorization and billing services, we call this provider a *virtual operator*[10] [62]. Thus, virtual operators do not need to have their own infrastructures. Typical virtual operators (VO) are e.g. 2G or 3G providers (because of their existent user database), miscellaneous resellers but also credit card issuers, banks, public remote authentication services, etc.[11]

Providers may (but are not required to) serve users for whom they are not home providers. Providers may propose access to services in their own and in other infrastructures (e.g. in the Internet or in user's home network). The necessary network interconnection can be based upon private infrastructure interconnections of several providers or it can be based on a public backbone like the Internet. This and other definitions e.g. service level agreements, price agreements, mutual agreements on user authorization in visited networks, etc. are subjects of so-called *roaming agreements* signed between the legal bodies representing the providers. Using these roaming agreements, providers can verify identities and profiles of visiting users whom we call *visitors*.

---

[8] Since users are the main focus of our work, we prefer this term to the synonymic *operator* which refers to the infrastructure.

[9] This is not limiting since any legal body can have multiple user assignments.

[10] This is used consistently to the original definition given in [62]. However, since in this special case no infrastructure exists, the actually "operated" entity is the user. This term is thus also consistent with our strictly user-oriented view.

[11] That underlines the fact that our model mainly requires the service contract as a means for a reliable user identification. Indeed, without *any* pre-established trust, no reliable billing is possible.

The users who do not have any verifiable service contracts may be treated as *guests*. Guests are users with special authorizations (profiles), locally and freely defined by any operator. These are thus local users and will not be treated differently in the following.

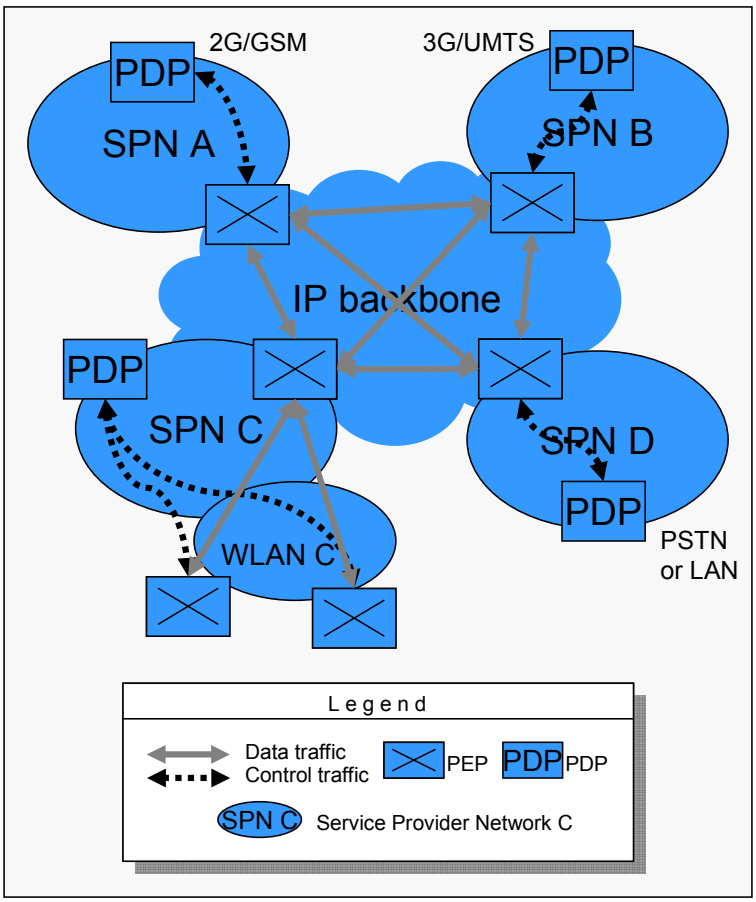


Figure V-1 Global system architecture

V.1.3. SPN Organization and Management

Management tasks in the SPN are carried out by the SPN owner, i.e. the provider. The actions are based on the management policies that reflect provider and user requirements. For this purpose, providers deploy policy decision points (PDP), i.e. logical entities capable of taking completely automated or assisted decisions based on the observed network situation and the defined policy. Policy enforcement points (PEP) are installed in the control equipment to enforce made decisions. In particular, PEPs are installed in the edge equipment.

SPNs are supposed to be trusted, non-public networks with appropriate protection measures. User traffic is to be strictly separated from the management traffic. The internal communications are IP-based. Inter-SPN management traffic can be protected by IPSec [76] or by using dedicated protected links (L2 VPN services, trusted sub-infrastructures, etc.)

Internal SPN architectures are addressed in more details in Chapter VI. We do not dictate the protocols and mechanisms regarding PEP, PDP, measurements, etc. Our main concern is to define architectures which do not impose any specific solutions. In a heterogeneous 4G system with its different providers (in terms of size, available resources, locality,

services, capital, etc.), this is an additional degree of freedom. Different approaches are principally suitable for management purposes such as proprietary console or Web-based management, SNMP [84], COPS [85], etc.

SNMP (Simple Network Management Protocol) [84] is an IP-based original IETF protocol. SNMP uses an agent-manager approach to read and to alter variables defined in the management information bases (MIB). For SNMP to be effective, the definition of consistent and sufficient MIB per used device is critical. COPS [85] is an IETF standard for exchanging network policies. COPS uses a mixed client-server/agent-manager approach. COPS is a centralized architecture consisting of a central policy decision point (PDP) and a set of policy enforcement points (PEP). This enables centralized QoS policy injections and their management on the PDP (on update push to the agents) and dynamic, load-dependent adjustments and traffic control based on the data from the PEPs (requests to the server), installed in e.g. network edges, routers, etc. The proliferation of COPS has been slowed down because COPS uses policy information bases (PIB) incompatible to SNMP MIBs and difficult to define.

V.2. System Access Model

V.2.1. User-centric Common Access Model

In this section, we define the access model which is used in our work. According to the ISO/OSI layers, there is a major difference between the network access (typically L2 or L3) and the service access. However, from the user's point of view, the network access is only the technological means of getting access to the services and not important as such. In fact, it represents for itself a service. In the general architecture outlined above, user *terminals* and access networks move relatively to each other, establishing connections to obtain access to the services in the SPNs. Our user-centric model accentuates this mobile/roaming service access, underlining that the serving SPN is not necessarily part of the infrastructure of user's home provider. This is illustrated in Figure V-2. User domain indicates the trusted domain of the user. In particular, this includes user's 4G terminal. The trusted domain of a provider includes the SPN and all ANs. User and home provider are expected to have a preliminary trust relationship (expressed and defined by the signed contract, see Section V.1.2). The visited and the home provider are expected to have a direct or indirect trust relationship, typically expressed by a roaming agreement (see Section V.1.2) or by an existing trust to a third trusted entity (e.g. a *broker*).

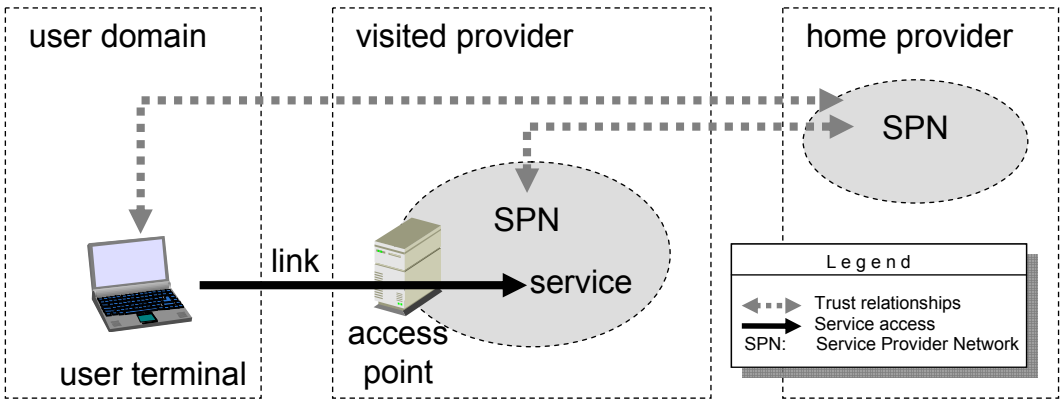


Figure V-2 Common access model

V.2.2. Service Access

The user accesses the service in the visited provider's SPN over the AN, AN's access points (AP) and the wireless link. The explicit service placement in the provider's SPN (s. Figure V-2) is not limiting. It is even necessary: if some service S_0 is located in the public domain (e.g. Internet) or in home provider's domain, the visited provider's SPN has to explicitly enable access to this services. This means that the visited provider has to provide a local service S_{local} which enables access over its infrastructure to S_0. Otherwise (i.e. if S_0 is within the user's domain), no network access is necessary – this case is out of the scope of the studied 4G problematic.

As in 2G and 3G, we consider the most general case of a multi-provider environment. Users access one of the available SPNs over one of the available ANs using their terminals. However, in previous generations the available services were only different because of provider's business considerations. Technologically all SPNs and ANs had (almost) the same possibilities. In the fourth generation, the services offered by providers can be different both in business and in technology dimensions. For instance, the offered services can be different in different SPNs, even if using the same AN technology. Some services might be only available over certain AN technologies. For example, if a user using a voice service over GSM wants to use a broadband data service, the user may be requested to access a locally available WLAN (operated by the same or a different provider). Also, the service continuity of this broadband data service may require provider changes, with or without access technology change.

Hence, if the requested service is not available in the serving SPN/AN combination, it can be used in any different SPN/AN combination to which the user has access authorization.

V.2.3. Roaming and Trust

In the context defined above, roaming access and mobility are critical issues since changing networks and allowing access to foreign users is of major interest for users (more suitable service, better service price, etc.) and providers (network usage means revenue) respectively.

The two existing trust relationships (user-home, home-visited) and a weighted transitivity are used to create a trust relationship between the user and the visited domain [86]:

$$\text{user} \sim_P \text{home AND home} \sim_{RA} \text{visited} \Rightarrow \text{user} \sim_{f(P,RA)} \text{visited},$$

where $\sim_X$ is the trust relationship for actions on objects defined in a set X, P – user's profile, RA – roaming agreement between home and visited providers. X is usually a list of authorizations. The function f combines the two authorization sets P and RA to a new authorization set describing user's authorizations in this particular SPN. Note that this model does not imply any type of identity verification (e.g. online/offline, symmetric/ asymmetric, security mechanisms, etc.)

Typically, data and control connections are necessary between the different providers.

V.2.4. Mobility Support

The mobility support is necessary for service continuity when changing the sessions, the ANs, the SPNs, etc. The handover can be classified depending on these changes as summarized in Table V-1. As can be seen, the smallest granularity of handovers here is the access network. We presume that handovers with a smaller impact area (intra-AN,

intra-AP) will be handled by the technology itself, i.e. by the 2nd ISO/OSI layer. A horizontal inter-AN handover can be managed by L2 or an IP micromobility suite [34], depending on the integrated mobility support and the size of the given network. In practice, we expect it to be managed by L2 whenever possible. This expectation is based on the observation of current standards. The L2 mobility support seems indispensable for limited range technologies since the overall coverage needs to be increased by adding access points. However, for a vertical inter-AN handover, L2 solutions can not be used because of the technological divergence. IP micromobility [34] is the only existing alternative so far providing a very fast and technology independent handover support. Although principally Mobile IP and higher layer solutions could also be used in this scope, this would be suboptimal since the higher level mechanisms result in higher handover latencies. With respect to the current very low proliferation of the IP micromobility in practice, this case represents a problem scenario.

Table V-1 Handover classification in the used system model

Type of handover		Direction	Impact	Layer possible	Suitable mechanisms	Expected practice
Same SPN	New AN with the same technology	horizontal	micro	L2, L3	L2 intern, IP μmob	L2 intern
	New AN with a different technology	vertical	micro	L3	IP μmob	No support
New SPN	New AN with the same technology	horizontal	macro	L3, L7	MIP, SIP	SIP
	New AN with a different technology	vertical	macro	L3, L7	MIP, SIP	SIP

An authority change occurs for a given session when users change the SPN. This authority change implies macromobility. Mobile IP (MIP) has been designed for macromobility support in IP networks. It could be used in that scope. However, the actual proliferation of MIP is low. Also, MIP is very infrastructure-centric. With the expected shift towards the user-centric design, we expect such cases to be managed on the session level, i.e. by an application level protocol. Mobile SIP [87] is a suitable solution that is currently rapidly evolving.

V.3. Logical Access Problem

Using the access model introduced in the previous sections, in this section we illustrate why the logical access to 4G services is not trivial. Different related underlying problems in security, signaling and service access areas are discussed in Section IV.4. Here we first define what we understand under the logical access. We then illustrate different issues that have to be resolved.

V.3.1. Definition

Prior to any service access, the user has to be able to communicate with service providers. In 4G, this requires the establishment of physical layer radio communication channels and some basic bidirectional communication capabilities at the data link layer. We call this procedure *medium access*. This is the task of the respective L1+L2 technology definitions. Note that this does not necessarily conform to the medium access definition of the respective wireless technology. For instance, in 802.11 medium access can include a device authentication part.

We define the *logical access* as the set of actions that have to be dynamically carried out in the three domains (user, visited and home providers) to provide a user with access to services.

In our definition we implicitly separate static radio technology and link establishment related parts from the dynamic access mechanisms including e.g. device configuration, network optimization, QoS, security, and other related measures. This dynamism is necessary since wireless technologies could use different dynamic parameters on the physical and on the data link layers. For instance, a CDMA-based technology implying per-user chipping sequences as a security mechanism is an example for such a case. Using the dynamism, we can distinguish between technologies that require a dynamic establishment of the chipping sequence and between technologies that use the same preconfigured chipping sequence in the network adapter (e.g. 802.11 with DSSS [14]). The first type of technologies has to be treated within the logical access problematic. The second type of technologies transparently uses the preconfigured sequence and no additional logics is required.

Hence, the logical network access typically includes functionalities from different layers. Moreover, we distinguish four main procedural groups:

1) security

2) configuration

3) service discovery

4) service access

For the reasons discussed above, all four groups include only dynamic mechanisms. The security group comprises all security related mechanisms (protocols, methods, etc.) necessary to access the AN's communication service (the link), the SPN and the service itself. In particular, it includes the identification of all involved entities, if necessary, and the correspondent credentials and trust management, i.e. in particular the user and profile management. Configuration group includes configuration tasks at different layers of the involved entities, necessary to communicate both on the control and data planes. Specifically, this group includes mechanisms for a dynamic control entity discovery, address assignment, etc. Service discovery regroups mechanisms necessary to discover the services proposed in the accessed SPN. Finally, the service access part includes all procedures and mechanisms implemented to access the discovered service. This includes the necessary negotiations, service access, service termination, etc. but also other topics such as provisions for billing (e.g. accounting, journaling, logs, etc.).

We can think of these procedures as generic signaling necessary for session opening at the session beginning. However, most of these procedures could also be required as in-session signaling. For instance, the service discovery should not be limited to the session start since services could be shut down or revived during an open session. Also, security group has a special meaning in the above list. While the configuration and service related groups can be seen as a logical sequence of tasks from the session opening to a successful service access, the security is somewhat orthogonal to these: security services can be required at any stage of the process.

V.3.2. Problem Statement

The user-oriented design and our user definition (see Section V.1.2) imply one contract per user with some, possibly virtual, provider. Given the limited coverage of certain technologies and different services to be provided over different technologies, roaming access seams an important feature. There is also a potential need for session mobility.

The latter two points presume the possibility to (seamlessly) roam into a different SPN/AN combination. This presumes at least a potentially global roaming support with service access and security provisions for the used access technologies. Herein, a distributed user management, i.e. inter-domain authorizations, user identification, corresponding profile definitions, etc. represent major technological challenges.

In a general case, a dynamic multi-layer reconfiguration of terminals and network equipment is difficult since all widely implemented reconfiguration mechanisms are technology-dependent (e.g. IP or L2 mechanisms). Yet, using an appropriate mechanism at every layer / in every access phase drastically increases the service access delay.

Different mechanisms for network service discovery have been proposed so far [88][89]. However, in the case of 4G, the service discovery mechanisms must be personalized. Depending on the authorizations of the particular roaming user, the possible services can be restricted. For security and privacy reasons, the service discovery should reflect this restriction. There are also some issues with layered ISO/OSI models: the services should be discovered very early. Yet, their placement in the SPN requires an accomplished, successful AN access and the configuration and activation of the IP access to the SPN. This is problematic, since users should be able to discover networks without accessing these on the data plane.

The logical access security problem is implied by the general problem of the heterogeneous security. All wireless technologies provide some security definitions. Such definitions typically include authentication, encryption and signing methods. These methods have different requirements in terms of security associations (see Section III.5.4). According to the user-centric design (see V.1.1 and V.2.1), user authentication (as opposed to device or host authentication) is used in our 4G model. To provide a reliable billing, the accessing devices should be bound to this initial user authentication. Thus, there is an urgent need for mechanisms capable of deriving the necessary security associations from the initial user authentication. Thus, we have an issue with the dynamic inter-layer security provision for the used technologies.

We discuss several specifics of these problems in the following. First we discuss different problems with the initial user access to a given SPN (roaming). Then, we illustrate user-related concerns in such a heterogeneous multi-technology environment discussing among others network and service detection and selection problems in our 4G vision.

V.4. Optimizations to User Roaming

V.4.1. Introduction

Roaming has been identified above as an essential service in the future 4G systems for both users and providers. Roaming as such is applicable when necessary trust relationships can be established (see Section V.2.3). However, this further technical means are required to enable a roaming service. In what follows, we identify some potential problems with 4G roaming and present our approach.

Without further technical means, a user can only access her home SPN, i.e. the infrastructure owned by the provider with whom the user has previously signed a service contract. Consistently with the system definition given in Section V.1.2, for each user, the home provider is exclusively responsible for:

- The definition and the maintenance of the associated user management infrastructure (databases, protocols, etc.)
- User management routine (user profile changes, creation and deletion, etc.)
- User authorizations and user group assignment (allowed services, their prices, maximum/minimum QoS, etc.)
- Home SPN and home AN access control equipment management (if any)

In a heterogeneous environment as it has been presented in Section V.1.2, roaming can occur between the networks of two distinct providers. Collaborating to extend their mutual potential coverage, these operators are likely to sign roaming agreements. In these roaming agreements, providers have to define contractual scope of their collaboration and the technical measures for its realization.

Managed independently and exclusively, the user profile databases of the two providers are likely to be incompatible in terms of the access protocols and of their content. Moreover, the access control equipment within the network infrastructure is exclusively controlled by the protocols and trusted entities of the owner, generally incompatible with the access control equipment and protocols used by the other provider. Ignoring other aspects of the roaming contract, the two providers have to agree upon several technical means for roaming user authorization, such as:

- Authorization data entities and exchange protocol
- Authorization data formats and meanings
- Consistent user naming schemes (to determine the appropriate roaming partner)
- Group/role meanings definitions in the visited network and their mappings

These technical means build the roaming support subsystem of each SPN and are situated in the control plane of the latter.

In the following sections we study the technical requirements of these roaming subsystems in light of our general requirements and complying to our 4G system vision.

V.4.2. Technical Requirements

From the system definition above and from practical considerations, we define the following requirements for the roaming subsystem.

1) Local user management

Consistent with the requirements of our system architecture, users profiles are treated explicitly by their providers (or by users themselves, within the defined constraints).

2) L2 access control

The security measures and the access control should be applied starting with the MAC layer. This is because of risks and the associated protection measures identified for the wireless systems in general. Protection on higher layers can be used additionally. However, since the latter is based on IP, it can be homogeneous and thus easier to specify.

3) Minimization of roaming access delay latency

According to our system model, the home and the visited domain communicate over global IP-based networks like the Internet. Such communication links are likely to exhibit Internet-level round trip times. For efficiency reasons, we want to minimize the number of messages to be exchanged on user access demand between the visited and the home domains. This aims to reduce the necessary signaling bandwidth between the two providers and the experienced access latency for the user. In the best case, the number of such messages can be zero meaning that all access procedures are carried out locally in the visited domain. This is e.g. the case for local users of the visited domain.

4) Rapid propagation of profile changes

The changes applied to the user profile by the home provider should become active as fast as possible in all possibly visited SPN/AN combinations. We call this parameter *system reactivity to profile changes*. Note however that roaming is defined as change of SPN/AN combinations at the session level (i.e. roaming alone does not provide session mobility). Thus, for the case of user roaming, the system reactivity to profile changes is optimal if the changes are enforced on the next session opening. For example, if a user account has been invalidated, the next access demand should be denied.

In the following section, we discuss the typical approaches and the existing solutions with respect to the requirements discussed above.

V.4.3. Existing Solutions and Associated Issues

Roaming needs to treat different interdependent aspects of the user-provider-provider collaboration. This can be subdivided in three major aspects:

- user to network access control methods including user, visited provider and corresponding roaming partner identification;
- definition of the local roaming subsystem (authentication, authorization and resource control subsystems for visitors);
- interconnection of the roaming subsystems of different providers (protocols, group mappings, etc.)

We discuss these aspects in the inverse order.

For the provider interconnection, the IP-based core network is used. That is because of our 4G vision introduced in V.1. The necessary interconnections can be private or public. Public interconnections need additional security measures, such as IPSec [76]. The necessary protocols depend on the chosen roaming subsystem.

The roaming subsystem should provide the user authentication and authorization and resource control possibilities. In the IP world, one well-known and broadly used example of such a system is the open IETF standards for Authentication, Authorization and Accounting (AAA) [90] including e.g. RADIUS or DIAMETER standard families. In the industry, RADIUS [91] has gained a broad acceptance since its introduction. However, RADIUS is based on a pure client-server communication model. This can be limiting in certain scenarios, e.g. changing policy enforcement. Also, since it is based on UDP, its support for the resource control is not sufficient. Some security issues have been reported for RADIUS architectures [92]. Generally, the usage of RADIUS can be problematic [91]. DIAMETER [93] is a relatively new Internet standard track that is designed to be backwards compatible to RADIUS. At the same time, DIAMETER architecture supports

additional communication models such as the agent-manager model. Using DIAMETER agents, this enables spontaneous decisions in the core network and their immediate enforcement at its edge (e.g. limiting actions based on diverse measurements). Since DIAMETER has been designed with RADIUS-problems in mind, it can typically mitigate the issues with the RADIUS architectures. Both RADIUS and DIAMETER are centralized AAA architectures featuring extensible protocols and explicit roaming support. These are also used in the GSM and 3G architectures. Principally, these can be used in our model for the following reasons:

- the AAA model explicitly targets usage scenarios with different access technologies;
- both protocols are kept independent of the access technology; the definition of the additional attribute value pairs (AVPs) enables protocol extensions supporting new technologies and their needs;
- by default, DIAMETER and RADIUS use IP as transporting network layer.

Finally, the first of the roaming aspects defined above depends on the used technology in the case of 4G (i.e. mutual user to network access control). Indeed, according to our second requirement defined in V.4.2, L2 access control has to be used. However, L2 access control is typically technology-bound meaning that the defined mechanisms are different from technology to technology (authentication, encryption and the resulting requirements).

In some cases, different requirements of the user to network access control can be rendered homogeneous by the roaming subsystem, which can hide the actual exchanges on the user link from the collaborating provider translating these into the defined exchange protocol. For instance, GSM/GPRS/EDGE uses an interesting abstraction mechanism to mitigate such problems: the trust definition in GSM/GPRS/EDGE includes the definition of virtual algorithms for authentication (A3) and key derivation (A8) and solely the encryption mechanism (A5) is to be implemented in the handset hardware. During the first user access, the roaming subsystem in the visited network (VN) can not verify these exchanges. It simply resends all data received from the user to the home network and vice versa waiting for HN to display the authentication result and to deliver the necessary derived key material in case of success.

Another interesting approach in this scope is the IEEE 802.1X architecture (Figure V-3). This architecture is an open framework for a port-based (L2) access control. From the security perspective, IEEE 802.1X provides an identification function without fixing a particular authentication mechanism. In this sense, IEEE 802.1X also represents an additional abstraction layer virtualizing the actual access methods similarly to GSM. It integrates with AAA architectures [90], which can be used in this scope to centralize profile, security and network management. The authentication method here is part of the user profile and can be dynamically negotiated between the user and the AAA server.

IEEE 802.1X explicitly mentions usage with RADIUS [80]. The APs use RADIUS protocol to contact the RADIUS server on user connect. On the user link, IEEE 802.1X defines an authentication payload transporter as opposed to authentication method. This payload transporter called EAPOL [80] transports IETF's Extensible Authentication Protocol (EAP) [94] frames in the 802 frames. The contacted AP reads the transported EAP frames and resends these to the AAA server. Thus, every authentication method in an EAP form can be principally used.

Several proposals and Internet standards for a EAP based authentication are available. Among those are EAP-MD5 [94] and EAP-TLS (Transport Layer Security in EAP) [95]. A lot of other EAP based authentication methods are currently work in progress at the

EAP WG of the Internet Engineering Task Force (IETF) like EAP-GSS (Kerberos over EAP), EAP-AKA (3G authentication in EAP), EAP-TTLS (Tunneled Transport Layer Security) or PEAP (Protected EAP) [96]. Obviously, the suitability of every given EAP method has to be carefully investigated for its use with a given technology. For the authentication purposes in e.g. 802.11 networks, a requirements document is being developed by the IETF and IEEE [97]. According to this draft, EAP-MD5 is not suitable but EAP-TLS is a good candidate.

In IEEE 802.1X, the APs act as blocking bridged with AAA-clients. They block the incoming data traffic from and to every wireless port prior to its opening. EAP frames are treated as special management frames. These are blindly resent to the AAA server using EAP-in-RADIUS attributes [98]. The port blockade persists until the arrival of the AAA-server's positive response message (i.e. `Access-Accept` message in RADIUS). The AP then grants access to the port where the EAP message came from. For example, in the WLAN case the ports are in fact logically assigned after the association.

A more complicated case of a non-local user access is carried out by proxying means defined in the AAA protocol [91]. This is shown in Figure V-3. In this case, the AAA server in the visited domain (AAAF) simply resends all arriving requests to the home AAA server (AAAH).

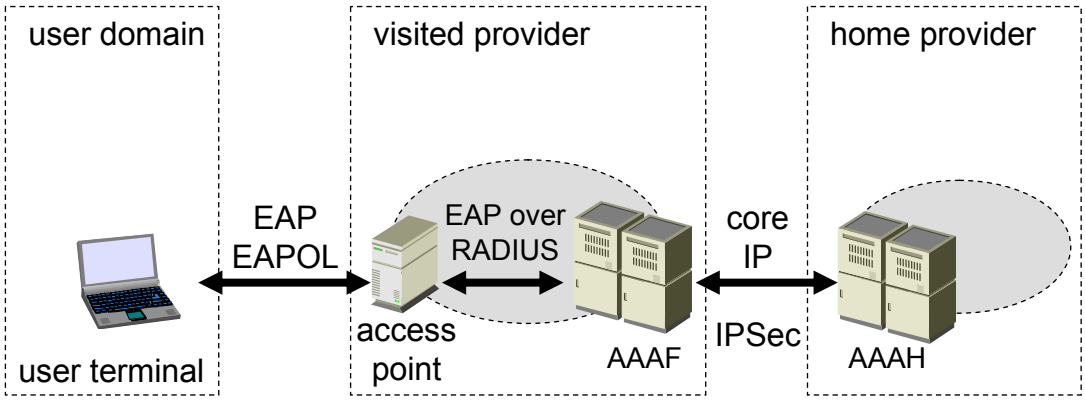


Figure V-3 General IEEE 802.1X architecture with roaming

GSM/GPRS/EDGE and IEEE 802.1X thus accomplish a potentially VN-independent user authentication with L2 access control and roaming support. Both solutions are used with a completely local user management (HLR in GSM/GPRS/EDGE, AAAH in 802.1X). Also, the system reactivity to profile changes is optimal since the managing domain can be contacted online on every user access to every SPN/AN combination. However, in both GSM/GPRS/EDGE and IEEE 802.1X roaming, the whole user authentication procedure is carried out between the home provider and the user. The visited network simply acts as a pass-through. These solutions are thus possible but not optimal. They do not minimize the roaming access latency (critical for a fast user reconnection) and the related number of messages exchanged over the backbone (referred to as *roaming cost*) as stated in our third requirement.

Studying WLAN roaming in 2001-2002 GET research project called AUTHENTIS [99], we have recognized that the system reactivity to profile changes and the roaming access latency are generally interdependent issues in the roaming systems. Tradeoffs are possible when adjusting these parameters.

For symmetric systems, a long time secret is pre-shared between the home and the user domains. There is no way for a third party (e.g. a visited domain) to know this secret.

Thus, the online identity verification is generally necessary. In some cases, the home domain can delegate the responsibility to a visited network by issuing limited verification tokens that can then be used locally. For asymmetric systems, it is often believed that an online verification is not strictly necessary. Indeed, given that the used certification authority of the home domain is known/accepted in the visited domain and the certification authority (CA) of the visited domain is accepted by the user, the visited domain and user can locally verify their mutual identities. This is e.g. the case when a commonly trusted CA is used by both home and visited domains. However, certificates can be revoked. This is especially true for user certificates: normally, users can not be prevented from quitting a subscription. If the expiration date in the certificate points to a date after the actual subscription end, the certificate has to be revoked. Thus generally, a visited network can only be sure of the validity of the user's identity if it can contact the home domain's certificate revocation list (CRL) i.e. a database of revoked certificates. Different methods can be used for these purposes, e.g. a complete or a differential CRL download, an online CRL checking protocol e.g. OCSP [100]. Hence, an online user validity check can not be avoided in this case either. We conclude that the mentioned relationship between the roaming cost and the system reactivity to changes is independent of the used user-home trust representation (symmetric or asymmetric).

It does however depend on the used authentication methods and the related protocols. This is where optimizations can be applied.

In the symmetric GSM/GPRS/EDGE systems, the roaming is optimized as follows. Upon a successful roaming access, the home network delivers N_t verification tokens to the visited network. These permit N_t successful consequent, completely local authentications between the visited provider network and the user. However, this optimization of roaming costs also cancels the optimality of the system reactivity to changes: indeed, an invalidated user can still access the same visited network until all issued tokens are used up and a new authentication to the home network has to be performed. In GSM, because of the wide area coverage of every SPN, provider changes are fairly rare. Also, the session duration in GSM is particularly long: once connected to the network, the user actually never disconnects. N_t can be thus chosen very low (in the order of 10). This optimization is thus successfully applied by most providers. Every provider can decide how high N_t will be, depending on the user and the visited network (typically subject of the roaming agreement).

In the 4G systems, provider changes are likely to occur more often because of the used short range technologies. Moreover, for security reasons, the session times are supposed to be shorter than in GSM. To achieve the same optimization degree, the number of issued tokens (N_t) thus has to be chosen considerably higher than in GSM. Also, since different access technologies can be used at the same time by every user, in the worst case scenario $k*N_t$ tokens have to be issued when k different technologies are used. This high number of possible sessions after the profile change significantly disturbs the system reactivity to profile changes and GSM's optimization can not be considered successful.

V.4.4. Main Idea: Authentication vs. Authorization

Another roaming optimization thus seems necessary. After a proper analysis of the issues discussed in the previous section, we come to the conclusion that the real reason for the home domain online check is not the verification of the user identity (authentication) but the verification of the current authorization set assigned to the user (user's profile). Indeed, even if user profile has been altered (e.g. deactivated), logically it does not mean that this user is not the same user. We treat that event as assigning the user an empty authorization set. The identity of the user and its validity are independent of this process.

Thus, in our vision, the user can be successfully identified without any implications on service access rights, etc. Instead, a respectively suitable authorization set (i.e. user profile) gives rights to access services.

Accordingly, we propose to logically separate user authentication and user authorization processes. We presume that the visited-to-home (VN-HN) bidirectional trust relationship and domain interconnections are already established to an agreed level (subject to a mutually signed roaming agreement). This trust relationship is used to provide a secure bidirectional communication channel between both involved providers, established independently. User authentication is a complicated security mechanism which typically involves several exchanges and complicated cryptographic treatments. Pre-established (bidirectional) trust relationships are necessary to be able to perform a (mutual) authentication. Compared to the latter, user authorization is a relatively simple process including a list transfer from one trusted entity (home) to another (visited). The main idea is to check the authorizations at every roaming user access since it is simple. The user is then authenticated locally. That minimizes both the roaming access latency for the user and the backbone usage for the providers, yet guaranteeing an optimal system reactivity to profile changes.

V.4.5. AUTHENTIS: First Step to a Virtual Home Network

Following our vision, the definition of the home domain slightly changes. Previously, the home domain implemented the whole access granting logics, mixing the user identity verification with the user validity verification. Now, we delegate the user verification to a visited domain. Thus, the visited network takes over typical home network functions. In that sense, this approach is the first limited step towards the virtualization of the home network definition. In effect, 4G users do not care what domain they connect to as long as it provides expected services at fair prices. However, the system security considerations imply the need to be sure that this is one of the trusted domains. In this approach, by delegating more and more functions to the trusted provider networks, the home network itself is now used solely as profile authority and as a common trusted entity of user and other providers.

To test our ideas with respect to their feasibility and practicability, we need a suitable technology with roaming support. As already mentioned before, the mixed AAA/802.1X has the interesting property of a L2 technology independent, EAP-based authentication method support. This neatly corresponds to our general vision of the security specifications in the next generation heterogeneous systems. The open standards and the availability of the AAA technology are another important factors. Finally, the main motivation is the fact that AAA is the technology of choice for fully converged IP based networks.

EAP methods play a crucial role in such access control architectures. Since its release in 1999, EAP-TLS [95] has been the only EAP method available as a released standard track that fulfills our security requirements (see Section IV.4.1): EAP-TLS provides both mutual strong authentication and dynamic session key derivation. It is thus interesting to take a closer look at this method that plays an essential role for today's practical WLAN security. EAP-TLS relies on the asymmetric cryptography for trust representation. User and user's home domain have a bidirectional trust relationship represented by a signed identity certificate with the corresponding private key and a self-signed CA certificate. Since the HN alone is responsible for user management, the certification authority (CA) and the associated CRL are situated within the home domain.

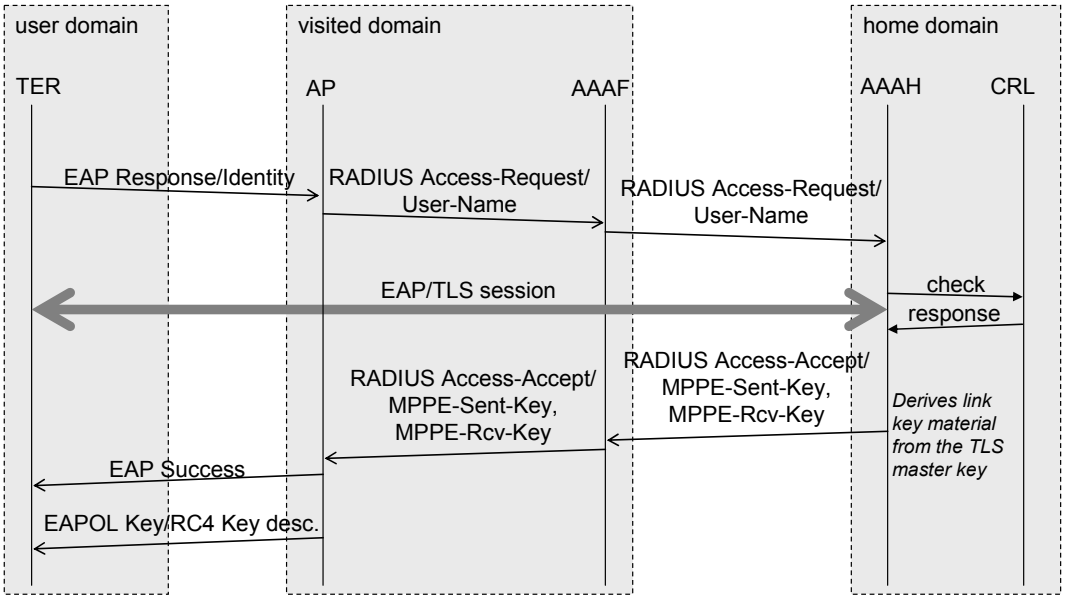


Figure V-4 Flowchart of a normal EAP-TLS operation with user roaming

The default exchanges of a user access to an AN implying 802.1X/AAA access control with EAP-TLS are represented in the flowchart in Figure V-4. As can be seen, the authentication method (EAP-TLS) is carried out between the user terminal TER and the home domain AAA server (AAAH). To resend the requests to the correct AAAH, the AAAF has to be able to determine the correct home network for every roaming user from the available user data. Except this decision, the user identity verification procedure is transparent for the visited network and its equipment (AP). The flowchart demonstrates a standard AAA operation mode with roaming. It shows that the AAA protocols naturally support an online user profile validity verification by relaying all incoming requests to the home server (proxying).

In case of a certificate-based trust representation, this can result in an unnecessary double verification of the user profile validity. First, the AAAH checks the validity of the account pointed to by the received unverified user identity (content of the AAA User-Name attribute). This requires a lookup operation in the associated AAA user data (not shown). The retrieved data record is checked for its validity (e.g. not expired, not deactivated, has right to access the requested service, etc.) The actual authentication procedure is extracted from the profile, as one of the authorizations. Then, within the TLS process, the certificate validity verification is triggered against the CRL that is part of a local PKI (represented here as a simple check/response exchange for simplicity). Logically however, the CRL certificate validity and the AAA database profile validity must be consistent. The opposite case is undefined within any reasonable security policy: if the profile is invalidated or not available, the certificate should not even be checked; if the certificate is invalid, the account should also be invalidated since no login is possible anyway[12]. Thus, this double checking does not provide any additional decision tokens for user access authorizations. It does however require an additional architectural, protocol and implementation logics complexity. Besides, it additionally delays the authentication.

[12] In some AAA server implementations, it is possible to successfully authenticate users having a valid certificate without any AAA DB entry. However, formally this case represents a special "all or nothing" authorization. For a complex security policy, a finer user authorization is required.

Hence, we discourage CRL checking when the AAA model is used with asymmetric trust systems. Instead, we suggest that AAA should be used to provide the authorization sets. This is more consistent with the AAA idea. Moreover, it simplifies the certificate structure. When additional authorizations need to be added to an existing user account, no recertification is necessary. The obligatory online check of the associated AAA databases is the central part of the AAA access method. It naturally replaces the CRL check: if the account is invalidated, the authentication method does not even need to be started, thus resulting in fast rejection without resource usage.

We use these observations and ideas in our test platform. In the following, we present some practical side issues and our implementation. We have proposed and implemented a roaming solution representing a compromise between the reactivity of the user management and the minimization of the roaming access latency [99][101].

V.4.6. The Implementation of Our Approach

The complete final architecture of the proposed solution is presented in Figure V-5. In some preliminary phase, the Domain A establishes trust relationships with users (subscriber contract) and with other providers, such as Domains B, C, etc. (roaming agreement). This trust is represented by identity certificates signed for the user and every trusted domain. Thus, each user has to store her own private key, the corresponding identity certificate and the certificate of the home CA (so called root certificate, typically self-signed). Each AAAF has to store one entry of the form <private key, identity certificate, CA root certificate> for every domain whose users can roam in its ANs.

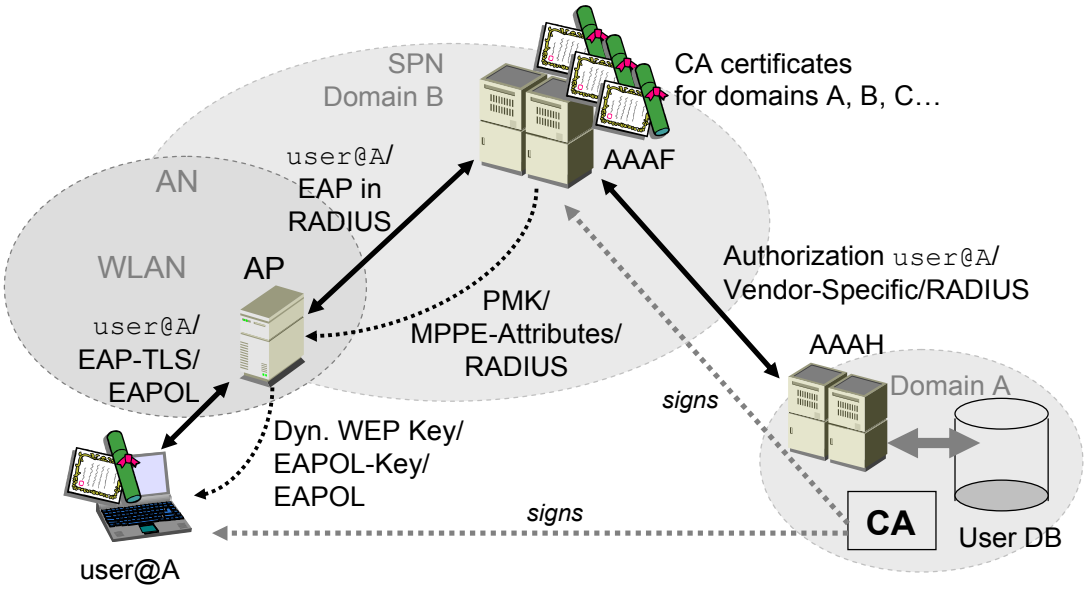


Figure V-5 AUTHENTIS final roaming architecture

When the user accesses a foreign domain B, the access point of the used AN (WLAN in our case) blocks the user's data traffic but forwards the user's EAP primitives to the trusted AAAF in the SPN. The trust between AAAF and the APs in the AN of the same provider is pre-established. It is usually represented by the trusted provider-own, non-public SPN infrastructure and, in RADIUS, by a pre-shared secret.

The arriving EAP frames and the AAA control data in the AAA requests are used by the AAAF for checks against the domain B policy. If the policy permits roaming for that

special access context, AAAF inspects the user name to determine the user's home domain. The attached database of the AAAF permits to map the realm part of the user name to the location of user's AAAH (IP address, port).

AAAF now requests authorization level for the user at the AAAH (identity transported in the AAA `User-Name` attribute). Technically, it removes the `EAP-Message` attribute from the incoming RADIUS request and forwards the message to the AAAH using its proxying mode. Without the EAP attributes, AAAH does not start authentication. AAAH replies with the list of the user's authorization levels (adapted user profile). We added a software module that uses a new vendor specific attribute to transport these data. With this method, profile existence, profile validity and user authorizations can be verified.

Once the authorization levels are verified and the user has the necessary rights to access this particular network the AAAF proceeds with the EAP-TLS authentication process with the user. For this purpose, AAAF chooses its identity certificate signed by the CA of Domain A. If the user authenticates successfully (i.e. presented certificate is valid, user possesses the corresponding private key and the value of the User-Name attribute corresponds to the identity in the certificate), the server issues the AAA `Access-Accept` message to the AP that has sent the initial request. Key material for the technology-dependent cryptographic suite on the link (e.g. encryption, signing, etc.) is added to this message. The AP uses this key material to derive the necessary key hierarchy for the link, applies the protection measures and opens the port for user data traffic. The AAA servers thus remain technology-agnostic. User TER can do the same locally on reception of success confirmation (`EAP Success`). Multicast keys have to be delivered from the AP to TER. The corresponding message exchanges are presented in Figure V-6.

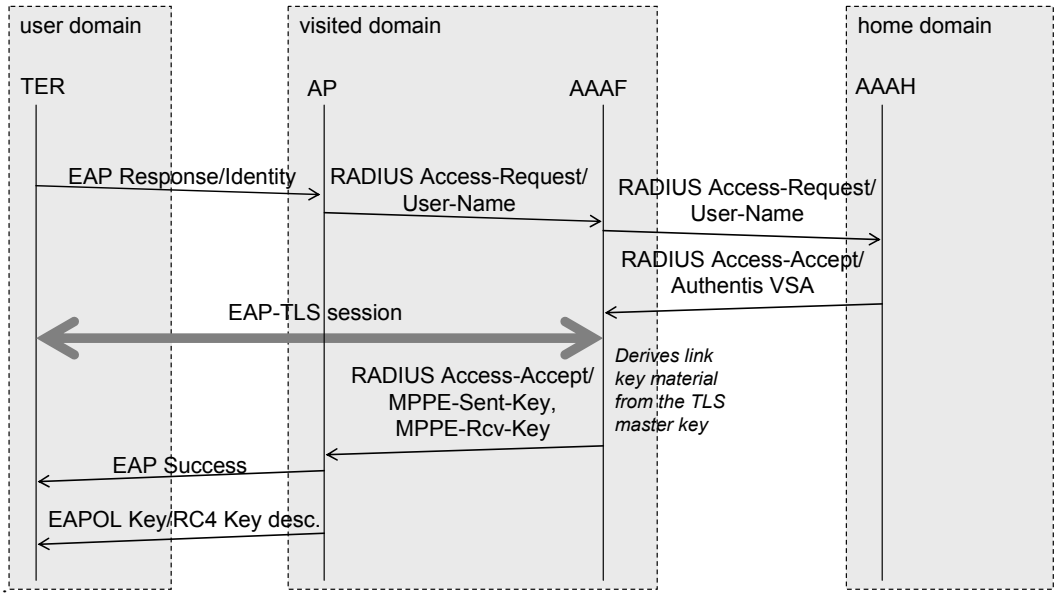


Figure V-6 User roaming exchanges in AUTHENTIS

V.4.7. AUTHENTIS platform

In our project work started in 2002, we have built a testbed implementing different standard case scenarios. The test platform setup itself follows the model presented in Figure V-3. Changing the respective behavior (logics) of the AAAF and AAAH and the EAP methods, we used this platform to evaluate user roaming with IEEE 802.1X/AAA

with different particular setups as mentioned in [99]. We namely implemented three different approaches representing all existing EAP standard tracks:

- first approach based on the symmetric EAP-MD5,
- second approach based on the asymmetric EAP-TLS using full forwarding to the home domain,
- third approach based on the asymmetric EAP-TLS using a common CA,
- finally, our approach as it has been presented before.

After different experiments and security audits with different tools in this real setup, we have come to the following practical conclusions:

1) Encourage 802.1X usage

We have evaluated alternative approaches such as PPPoL over 802.11 and L3 VPN usage for user confidentiality. Compared to the PPPoL approach, 802.1X turns out to be more efficient: although PPP [102] uses the same access model and can achieve the same security goals it is too complicated in the WLAN case. A big part of PPP is dedicated to the usage over connection-oriented links. 802.11 is packet-oriented and does not require PPP packet service establishment. In turn, a L3 VPN technology can not provide infrastructure protection for the deployed WLANs. Indeed, the access to the actual WLAN infrastructure remains open per default. Hence, we have concluded that 802.1X architecture is the best available method to enforce both infrastructure and user security in the existing 802 networks. Using a suitable EAP method and fast rekeying, 802.1X avoids the particular problems with 802.11 mentioned in Appendix A. Additionally, 802.1X pushes the system authentication to the user level. This in turn enables the definition of network-wide security policies defining exact user authorizations to different 802.11 services. Finally, 802.1X seamlessly integrates with the popular AAA systems and thus enables e.g. single sign on services (SSO) and accounting implementations.

Today, 802.1X is the chosen reference architecture for both WPA and IEEE 802.11i [81] standards. Also, 802.1X is the 3GPP candidate for future UMTS/WLAN roaming [103].

2) Studied and discouraged the use of EAP-MD5

For various reasons (no mutual authentication, doubtful MD5 strength [104], PRNG strength, clear text password, dictionary attacks, etc.) cited in [99], we have explicitly discouraged the use of the EAP-MD5 standard track with wireless networks. This is the accepted best practice today (e.g. the EAP-MD5 support for WLANs has been removed from Windows XP in 2003). The MD5 algorithm is considered broken since collisions found in 2004 [105]. Its usage is generally discouraged now.

3) Encouraged EAP-TLS usage with dynamic session keys and mutual authentication

The EAP-TLS chosen as candidate for our implementation is the technology of choice for the expected 802.11i security enhancement standard [81] for 802.11 networks.

4) Encouraged IPSec usage for AAA roaming over public networks

According to the RADIUS protocol, data communication between the server and the client does not transport any passwords (in particular, the `User-Password` attribute) in clear. But, RADIUS dictates the usage of MD5-based method for hiding the password value. There are some serious known security issues, e.g. described in [104][113]

showing the possibility for successful attacks against the used method in particular. Thus, as the RADIUS client-server security is considered weak, we conclude that the RADIUS traffic is not sufficiently secured to be transported "as is" over public networks, e.g. over the Internet. However, this is exactly what will happen in the most general case of roaming. So, in order to secure the server interconnections we strongly encourage the usage of IPSec on the underlying network layer[13]. The total number *i* of server interconnections for a configuration with N domains is:

$$i = \frac{N\ (N\ -\ 1)}{2},$$

which results in 3 interconnections in the current AUTHENTIS platform regrouping three French telecommunications universities (*N=3*). With this low number in our setup, we use IPSec in the following modes:

- Shared secret mode, hence the total of *i* shared secrets. The re-usage of network internal RADIUS client-server shared secrets is categorically discouraged.
- Transport mode, given that no tunneling is to be done and the transport mode is much more efficient.

As has been stated before, the AAAF has to be able to determine the correct roaming agreement partner. A proper user naming scheme in this context is important. Consistent with the current AAA best practice, the user name is used for this purpose. We adopt a network access identifier (NAI) [106] format. In our case, it directly corresponds to the email-address of the user, i.e. the complete user name is the same as the user's usual email address. This format is `user@realm` where:

- `user`: the user identifier, treated by the responsible RADIUS server; in our case it is the user's usual ID.
- `realm`: the identifier of the user's home network; in our case it is the mail-address-domain part of the respective organization.

Every AAAF not configured as responsible for the requested realm has to find the AAA server responsible for this realm according to its configuration. Then, it resends the request acting as an AAA client.

V.4.8. Comparison and Results

The results in comparison to other 802.1X roaming systems are summarized in Table V-2. It shows the used trust representation (user-visited and visited-home), the number of messages exchanged over the backbone on each roaming user access, the optimality of the system reactivity to profile changes, the requirement for online CRL check and the number of secrets in a system with `k` users and `n` SPNs. GSM/GPRS/EDGE roaming is not directly comparable to these systems because of its different complexity. We add it for reference only.

As can be seen, our system minimizes the number of delay-prone inter-SPN messages however maintaining the system reactivity to changes. E.g. GSM/GPRS/EDGE roaming uses its own optimization for the same matter, however it does not provide the optimal system reactivity. Compared to other 802.1X/AAA roaming solutions, our system uses even less inter-SPN exchanges than a very simple, symmetric EAP-MD5 challenge response mechanism. Our system is a security system based on the asymmetric cryptography both for user and domain trust representations (we briefly call it asymmetric

[13] TLS can not be used because RADIUS is a UDP based protocol.

system in following). A normal forwarding of EAP-TLS results in a dozen of inter-SPN exchanges on the backbone (the exact number depends on the fragmentation results, certificate size, etc). Our approach thus achieves a considerable improvement here. Internet level RTT of about 100-250ms is expected between two arbitrary SPNs. Minimizing the message number from 12 to 2 can help to dramatically decrease the network access delay.

Table V-2 Comparison of our proposal to other roaming systems for n SPNs and k users

Roaming architecture	trust base	inter-SPN Messages	optimal reactivity	CRL check required	Number of secrets
GSM/GPRS/EDGE roaming	symm.	0 (2)	no	no	n(n-1)/2 + k (no certificates)
802.1X/AAA with full EAP-MD5 forwarding	symm.	4	yes	no	n(n-1)/2 + k (no certificates)
802.1X/AAA with EAP-TLS without forwarding (global PKI)	asymm.	0	no	yes	n+k+1[14] (2k+2n+1 certs)
802.1X/AAA with full EAP-TLS forwarding	asymm./symm.	~12	yes	no	n(n-1)/2 + k (2k + 2n certs)
AUTHENTIS solution	asymm.	2	yes	no	$k + 2n^2$ ($2k + 2n^2$ certs)

Further optimization is possible in our solution in some cases: if the user authentication method is pre-defined by some means, the local user authentication (EAP-TLS) and the user authorization can be started simultaneously without any impact on the system reactivity to changes or on security. The AAAF has to delay the sending of Access-Accept until both following conditions are true: user has successfully finished the TLS authentication; user authorization reply from the user's home domain is positive. Thus, both delays (local and remote check) do not necessarily sum up in our approach. This optimization can not be applied in "full forwarding" approaches.

In asymmetric systems, presuming a full trust into some entity (i.e. a system-global PKI), one could effectively eliminate the need for full forwarding since users could always be verified locally. Besides the fact that this assumption seems unrealistic, it also violates our requirement for completely local management. But even ignoring these issues, the resulting solution would not provide an optimal system reactivity to profile changes (since changes would have to propagate to the system-global PKI). Finally, CRL checking at this global PKI would still be required, effectively resulting in a number of backbone message exchanges (exact number depends on the used CRL-check method).

Our solution reduces the complexity of the SPN architecture by eliminating the need for an additional CRL repository. It thus avoids the associated access, maintenance and administration expenditures.

Discussing the security of the inter domain connections we note that the only sensitive information exchanged in the inter-SPN messages used in our approach is the user name and the user's authorization levels. Even if this information was sniffed and decrypted, no successful access would be possible. In our platform, the messages exchanged over the backbone still pass through the secured IPSec channel. Given the relative insensitivity of the transported data, this can be considered sufficiently secure.

In terms of the number of secrets for a system with `n` domains and `k` users, `k` secrets are necessary for `k` users, independently of the used trust representation and roaming system. In the asymmetric systems, users have to additionally store their identity and CA root

[14] One additional (private key, certificate) pair is necessary for the global CA

certificates (both of which are, however, public). On the contrary, the number of domain secrets depends on the used trust relationships and the authentication mechanisms to users and other domains. In purely symmetric systems (both user authentication and inter-domain trust are based on pre-shared secrets) common secrets are necessary with every existing domain. This results in a total amount of $O(n^2+k)$ secrets for n domains with k users. For partially asymmetric systems like EAP-TLS with full forwarding (i.e. user-domain trust is based on asymmetric, inter-domain trust on symmetric cryptography), the overall number of secrets is thus the same. For fully asymmetric systems like EAP-TLS with a global PKI (i.e. inter-domain and user-domain trust are based on the asymmetric cryptography), every domain has to store its own private key and the public identity and CA root certificates. Additionally, the global PKI has to store the private key of the root CA. By fully localizing user management, our approach is fully asymmetric. In our approach, every domain needs an identification vector of the form <private key, identity certificate, root CA> per existing domain (including itself). Thus, every domain has n such vectors, resulting in n secrets and $2n$ certificates per domain. The secret storage complexity of our approach is thus also in $O(n^2+k)$ (i.e. asymptotically not worth than standard EAP-TLS roaming approach). Note however that the minimization of this complexity was out of scope of this particular work. In our practical approach, we tried to remain compatible with the existing standard TLS mechanisms and implementations. Minimization techniques known from the web technology can be applied. For instance, multiple independent CAs can be trusted by users and providers (replacing the global CA), hierarchical certification can be used, etc. However, such changes could require changes in the user authentication method implementations.

Another point is that we currently do not take into account any billing issues. That means that accounting is supported but not optimized in our system. We use AAA accounting model in its standard way. Accounting data are generated by the access controllers (access devices) and sent to the AAAF. Using these data, AAAF builds a session database. This database can be used for later billing. However, this is a multiparty payment scenario and as such a complex problem [72]. Since it also includes inter-domain exchanges, it can be related to the roaming problematic. However, typically, the accounting report delivery is irrelevant to system performance (e.g. it can be sent after the user has finished the performance-critical reconnection).

We continued the development of our test platform presented in the previous sections. Our experiences have helped us building the integrated services WLAN, which is today operational at the ENST Paris. We also pursued the infrastructure virtualization paradigm in further research projects. This and the resulting architectures are presented later in this document.

V.5. Virtualization of Signaling

V.5.1. Problem Statement

Diverse problems with the All-IP 4G approach are discussed in Section IV.4. Although IP can be used as a pervasive data transport mechanism, main problem areas are heterogeneous security, mobility management and user-QoS.

We see all these problems as part of the more general problem of a sufficient control plane definition. From this point of view, the protocols for the security exchanges, mobility registrations, QoS negotiations, etc. are part of signaling. Signaling protocols are thus necessary to be able to transport the exchanged control messages.

While IP is to be used as a generic mechanism for data transport, the signaling is used for the orthogonal purpose of enabling such transport. That includes system configuration, user registration, user authentication, parameter negotiation and other functions. Signaling is thus used to be able to transport the data in the appropriate way.

In the All-IP solutions, IP is the common data transport. Thus, logically the main function of signaling must be to enable IP usage. This includes capability considerations (can it be done?), organizational considerations (how can it be done?), and security considerations (is it allowed to be done between these entities?). Thus, we come to a logical chicken and egg dilemma: how can signaling be based on IP if it is used to enable IP in the first place?

Of course, in practice such problems are never difficult. They are usually resolved by loosening the constraints. For instance, some limited initial IP exchanges can be allowed regardless of the capability, configuration and security considerations. That typically results in initial multicast messages sent to some well-known IP address (e.g. limited scope broadcast address). Such mechanisms can be both passive (wait for a message) and active (force sending a message), which is merely a question of point of view. The widely-used Dynamic Host Configuration Protocol (DHCP) [107] or Agent Advertisement and Agent Solicitation mechanisms in Mobile IP [31] are examples of such a solution. These mechanisms work. The doubts about such mechanisms lie elsewhere, typically in their security, their purity and their efficiency.

The purity question arises when one tries to implement such solutions. It turns out that it is virtually impossible to remain independent of the underlying layers: indeed, ideally the IP layer needs an event trigger to be able to send an advertisement/solicitation message since otherwise it has to send such messages periodically, independently of their sense. That is amplified by the demand for more efficiency. Periodic messages are not only unnecessary, they also imply a tradeoff between the associated latency and resource wasting. Unfortunately, the definition of such events triggers results in a cross-layer design: the event triggers have to be defined in the underlying layers according to the needs of the IP layer and the respectively provided services and associated expectations. The trigger discussion gained momentum during the efficiency discussions of Mobile IP handovers in the IETF. The need for different L2 triggers has been recognized. Nevertheless, the IETF has mostly failed to impose such triggers for Mobile IP. Thus, today Mobile IP principally relies on periodic advertisements.

Thus, from the capability and configuration point of view, we have to admit that IP-based signaling partly relies on the L2 technology and thus requires some kind of homogenization efforts in the underlying layers. The situation is even worse for user QoS-classes. What can be done, if the IP-defined QoS classes are not supported in the layers below? How can a reasonable mapping be defined in light of complete heterogeneity? However, if the homogenization in the underlying layers is necessary anyway, what is the presumed advantage of the IP signaling?

Another concern is the security consideration. The loosening of constraints would result in uncontrolled, pre-authenticated exchanges on the IP layer. This needs holes in the access controllers (firewalls, etc.) and permits IP datagrams to be sent from arbitrary hosts to provider's control plane. However, IP is usually implemented in software and is believed to be much more vulnerable. Diverse successful attacks have been implemented against IP systems (Ping of Death attack, ICMP-Redirect attacks against VPN systems, SYN flooding, Denial of Service attacks, etc.) and new attacks against diverse operation systems are reported almost every day [108]. Indeed, the attacks based on the Internet technology are the main current security concern. That explains why the telecommunications providers mistrust the IP technology and hesitate in deploying exclusively IP-based security solutions. Instead, they usually rely on their own

mechanisms. That is also one of the reasons why we demand a reliable L2 access control in our security requirements in Section IV.4.1. Facing such L2 access control, the IP-based signaling has no chance: the complete L2 access procedure (physical and logical) has to be finished before any IP datagram can be transmitted. The corresponding delays thus sum up the worsening of the performance. Moreover, once again we notice that, at least in our view, a part of signaling is outsourced from IP to non-IP mechanisms.

For the reasons given above, we do not believe that sufficient, consistent and efficient exclusively IP-based signaling is possible. That position is backed up by the observations of the current situation given in the Sections IV.3 and IV.4.

Hence, in 4G, a generic non-IP signaling part seems to be required. Following the above explanations, this signaling method has to be situated below the IP layer in the ISO/OSI stack (since everything above IP will use IP as the unifying transport). Non-IP signaling is already in use in provider networks. In particular, that includes all integrated logical access procedures of the used L2 technology (L2 mobility management, L2 security solutions, etc.). However, such lower layer definitions are the main reason for the system heterogeneity. Definitions of the logical access methods, configuration, security, service discovery, QoS classes and various miscellaneous parameters are currently precisely tailored to the requirements of the used L2 layer. However, in light of such widely-used, generic, lower-than-IP solutions like MPLS [73], we think that there is some hope for a provision of a generic lower layer signaling channel for 4G.

V.5.2. Towards 4G Signaling: a Virtual L2 Signaling Transporter

Our 4G visions follows the All-IP solution for data transport. However, on the user-link it pairs this All-IP approach with a pre-IP signaling. The corresponding signaling is used in the control plane of the SPNs and in the terminal equipment in the user domain to obtain access to services. Thus, the terminal principally has two logical channels connecting it to every SPN: the signaling channel for parameter negotiation and a data channel for data transport to the IP based core backbone. This is illustrated in Figure V-7. As can be seen, we modify the introduced access model accordingly (see Section V.2.1). The new access model introduces two separate planes in each SPN. The data plane, accessible over the IP-based data link, implements the data access to services. The control plane, accessible over the signaling, implements the control elements for such access. Both access methods can be prolonged to other domains (e.g. to the home domain). If such prolongation is required, IP-based protocols can be used to exchange the signaling traffic between the domains over the IP-based core network. Thus, our change is limited to the SPNs and does not imply anything else than IP support in the connecting core network. The trust relationships remain untouched by this change.

In the classification of the possible approaches to 4G given in Section IV.3.2, our proposal represents a common, technology-independent access protocol. The practical path to the realization of that proposal is the extraction of the user-related access procedures from the respective access definitions of the used access network technology. According to the definitions given in Section V.3, this is thus equivalent to the separation of the logical access measures from the respective physical access measures.

To do so, we need to analyze the existing technologies and the user needs. We could externalize the common user-related part of the access phase and define these procedures independently in an extern, common method. This method could then be used with every technology. However, the related procedures are quite different and the homogenization of these would require impracticable changes to all technologies. Instead, we propose to leave the methods where they are and to externalize the decision making.

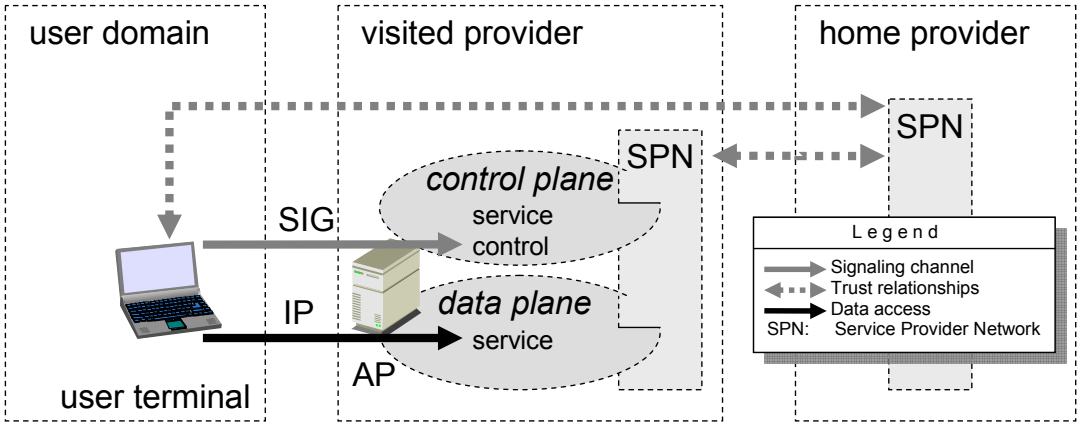


Figure V-7 Modified access model with two distinct access channels to the SPN

This needs some support for every necessary method in the SPN. Since the ANs and the SPN are under the same authority, the SPN can be prepared to serve all connected ANs. What we need to do is to separate the actual method from its transporter. The AN technology has to be able to establish some basic link and then to transport special management frames between the user equipment and the access devices in a distinct manner. This distinction would enable the access device to give these a special treatment, different than the usual data frames. Then, these signaling frames could be delivered over the AN into the SPN. The SPN could make the decisions according to its policies. The control points in the SPN control plane can command AN's access devices appropriately.

We want to illustrate this approach with a fictive example. Let the access technology T_1 define and support a QoS class Q_1. The necessary negotiation procedures vary from the access network to access network in terms of logics, formats, etc. Today, if a user needs to access a service through the access technology T_1 with quality Q_1, the user would use the procedures defined in T_1 to do so. At some point, the solicited access device needs decisions on authorization, resource availability, etc. These can be made locally, however in managed networks, these decisions need to be made in accordance with the SPN control plane. Hence, the access device has to contact the SPN, to demand a policy-based decision or a policy for that special user request. For that reason, it might use a generic procedure G_1 that transports the parameters from the user's request, the device's own state, etc. to an SPN control point. The control point meets the decision (or finds a correct policy) and gives the decision (policy) to the access device. Logically, this is equivalent to the situation where the user asks directly at the SPN's control point. The access device does not have to support the procedures defined in T_1. Using our generic transporter approach, the device acts like a medium: it simply resends every request to the control point, using the generic G_1. That is an improvement: first, the access device logics is simpler since it does not need to understand QoS related contents. The resending involves blind copying from the generic signaling transporter on the user link to G_1 on the internal link. Moreover, if the user now wants to use Q_2 through T_2, the procedure on the user-link remains the same. Thus, the user device logics can also be simpler.

Hence, this approach virtualizes the signaling, pushing all user-related procedures to the user-level. Bringing these decisions into the common control plane, it enables policy-based layer-spanning decisions that can take into account the parameters from different layers. The user specific access part has to be extracted from the generic frame transport of the respective technology. Further requirements of this transporter are discussed in the next session.

V.5.3. Requirements on the Common 4G Signaling

The transporter frame should be a simple container with a minimal overhead that can be transported by every technology without any or only with minor changes. Since signaling exchanges could be initiated both from the user domain and from the visited domain, the used mechanism should be based on a suitable communications model, allowing messages to be initiated and received by both communication parties on the user link.

Pre-authenticated Signaling

In particular, this mechanism should allow for pre-authenticated signaling. This is important to be able to fulfill the requirements on network access explained in IV.4.1 (keyword: network selection).

Since 4G is user-oriented, no limitations should be introduced here in terms of what should be allowed as preliminary network information. Network discovery can be extended to service discovery, which could include service endpoint discovery, available service QoS, service parameters and service prices for that particular user.

The pre-authenticated signaling is also important since it can help to prepare a handover. This allows for a smoother change. Handover latency is a critical issue for service continuity and perceived connection quality. Latencies of 200ms and more can be too high for certain service classes. Normally, the authentication procedure has to be started first, augmenting the overall handover delay. This can be avoided by inverting the situation: the necessary preparation (e.g. context transfers, QoS reservations, route changes in the SPN, etc.) can be triggered before the authentication. The authentication could be used after the handover is performed. That makes more room for optimizations of the latencies: e.g. some procedures in the SPN and the authentication could run at the same time. Alternatively, the authentication could run after the context transfer thus becoming completely local and thus faster.

Pre-authenticated signaling does not necessarily mean insecure signaling. In certain architectures signaling messages could be e.g. cryptographically signed and encrypted.

Authentication Signaling

The signaling transporter should also be able to transport the authentication signaling, i.e. all messages exchanged during the authentication. That does not change the authentication method itself. Its data are simply transported over another medium. The signaling mechanism should allow authentication method negotiation to provide more flexibility.

Post-authentication (in-session) Signaling

The same channel should enable the in-session signaling (i.e. post-authentication). That is because of the presumed dynamics of the 4G SPNs. If the service environment for the user changes (e.g. because of the SPN policy, or available resources or determined user credit, etc.), that should be shown to the user. This is usually done by dynamic service discovery mechanisms.

Another suitable candidate for the in-session signaling is the QoS re-negotiation. This could be important e.g. because the user has finished the voice call and now wants to transmit data traffic with more burstiness but without guarantees.

Mobility related state changes can occur in-session, following certain events. For instance, registration time could expire. That triggers re-registration exchanges without terminating the session. Paging is a mobility-related mechanism that can occur in-session

if the mobile changes the paging area. Similarly, location information should be carried over some channel.

Security functions could also require in-session signaling. A simple example is the re-authentication at the session time limit, to avoid expiration and disconnection. However, other examples are valid too, e.g. re-keying has to be enforced on both sides.

V.5.4. Possible Implementations

Different implementation possibilities exist to realize this idea. For diverse reasons discussed above, such channel can not be IP-based. Thus, the implementation of the actual signaling exchanges has to be in L2 or between L2 and L3 (L2.5). The user-signaling methods are implemented classically in the application layer.

New Equipment-Independent Signaling Channel

Principally, our idea can be implemented by defining a new generic transporter frame. The standard approach is the definition of an abstract Code-Type-Length header and the addition of appropriate code and type meanings. Then, every technology specification has to be slightly changed to allow the distinct transport of such frames. The access devices have to treat the L2 frames transporting the generic signaling transporter frame in a special way. However, the latter two changes are minimal and do not represent any technological challenges. All existing L2 frames allow some type of frame marking (e.g. marking as management frames).

Generalization of the EAP Approach

Our proposal can also be implemented by reusing the existing technologies. This could provide a backward compatibility to some extent. However, reusing existing technologies is also bound to some constraints.

A closer glance on the IETF's EAP [94] technology used for port-based access control reveals the proximity of the EAP idea to our proposal. Originating from the dial-up access control architectures as an alternative to its predecessors PAP [91] and CHAP [109], EAP can currently be used over a panoply of transport protocols including PPP [102], various AAA protocols, IP, UDP, EAPOL [80] and finally EAP itself. Because of this, EAP can be directly used as a user authentication protocol in

- 2G, 3G, xDSL, dial-up and virtual private networks (over PPP),
- in wired and wireless Ethernets (over EAPOL)
- and on higher layers using either IP or UDP or AAA as transport protocols..

EAP thus covers a vast variety of the popular user equipment. Since its introduction for 802, EAP has gained a tremendous popularity. Today, EAP-based authentication methods comprise certificate and password based authentication but also permit the usage of GSM SIM cards [103], SecureID tokens etc.

From our point of view, the EAP approach features the virtualization of the user authentication, pushing the actual authentication method away from the access device to the network core. It thus features separation of authentication/authorization that play a central role in the session admission decision and the link encryption that is merely one of the parameters in the link establishment procedure between the access device and the user terminal. Herein, the decisions are met at the *authentication server*. These decisions are enforced at the port controller, also called *authenticator* for the requesting entity.

Currently, EAP is limited to authentication purposes. The basic idea and the related architectures are however applicable for the implementation of our idea. Moreover, the authentication is an indispensable part of the user-network signaling. Given that reliable authentication protocols are difficult to design and using the EAP independence of the used medium, we could use EAP as the common access protocol in 4G.

Comparison

The EAP protocol itself has been recently revised at the IETF [94]. Because of its popularity in the 802.11 WLANs (over IEEE 802.1X), a lot of new EAP authentication methods have been submitted for standardization. However, at this moment, only two EAP methods have reached the status of IETF's standard tracks. IETF's EAP Working Group is currently working on the state machine definitions for EAP and is elaborating requirements on EAP-based authentication methods. The most recent EAP RFC explicitly discourages the usage of EAP for other than authentication purposes [94]. This is not because of our intentions. The problem is proprietary implementations using EAP for different, non generic purposes in some networks. We think that a proposition of a standard method could constructively discourage such usage.

While a completely new common L2 protocol would provide more flexibility both for our research and development (no constraints within the existing implementations and standards) and for our ideas (no organizational, political, etc. resistance of standardization bodies), EAP provides an immediate compatibility with the deployed equipment. For us, this is the most important point. However, we also inherit the burden of the deficiencies of EAP as discussed hereafter.

EAP follows a limiting client-server communication paradigm. All requests have to originate from the access controller (access point). The access controller issues the requests upon some events (e.g. L2 events like new connection, etc. which it obtains from its internal management plane). The user device replies to these requests. Thus, it is difficult to trigger EAP exchanges from the user device within EAP. Usually, it is done by some L2 specific method, like `EAPOL-Start` message [80]. This is of course a serious constraint since it reestablishes a (logical) L2 dependency.

Another limitation is that EAP does not currently support fragmentation. That effectively limits the PDU to the MAC frame size minus headers. IETF's EAP WG suggests to not presume EAP PDUs to be more than 1000 bytes. From our point of view, this limitation can be easily accepted when a prudent protocol design is applied.

Since EAP is already accepted by a variety of access controllers, EAP can be principally used as a generic user-dependent signaling protocol in the deployed networks. That has encouraged us to evaluate its capabilities by implementing our ideas in real networks. In the following, we describe our implementation.

V.5.5. Proposed Solution: EAP-SIG

With this motivation, we analyze the EAP suitability as a generic signaling transporter. As a standard example scenario, we use a 802.11 WLAN with 802.1X access control. This uses EAP between the supplicant (i.e. the user terminal) and the authentication server in the network core. The EAP traffic changes the transporter at the authenticator implemented by the access point.

In this example scenario, we first identify some potential problem scenarios. We then show our implementation. Finally, we show how the presented problem areas are resolved in our view.

Problem Areas

1) Common Access Protocol

The service access problem in a heterogeneous environment can be solved by using some higher level protocol (e.g. IP because of its pervasiveness) as a common access protocol. However, in presence of the aforementioned L2 access control this solution is not very efficient since it merely doubles the access control measures and the associated delays. Besides, a higher level access method has to additionally perform an access router discovery (usually involving broadcast) and a higher level configuration.

Conversely, leaving the access control untouched (i.e. limited to the respective link layer access control) lacks integration, necessary for a consistent 4G access.

2) Network Discovery

The roaming user faces the problem of choice between different available access networks. One of the problems is the selection of a network having a roaming agreement with the user's virtual operator [62]. However, even supposing that all users can freely roam in all networks, the problem of choice persists. If presented with two available networks of the same technology, how can a user decide which network to choose? Natively, e.g. 802.11 only delivers a SSID – an abstract network identifier without any guarantees. The user thus needs to fetch some network information before she finally accesses the network. Such information could include e.g. the prices and the available services for this user (if an authentication is already possible). Practically, this could be easily achieved by the means of an existing service discovery protocol [88][89][110]. However, as data exchange protocols these can not be currently used prior to the successful L2 network access.

Using these protocols after the L2 network access presents a twofold problem. Firstly, this is much less efficient since the actual L2 network access procedure has to be completed and higher level configuration and service discovery procedures have to be accomplished prior to actual discovery protocols. Additionally, higher layer service discovery unavoidably involves broadcast messaging or a station pre-configuration. Second, in the commercial networks the user could be (unfairly) billed before being able to choose (*pay-before-choose* problem [111]) if the accounting is activated immediately after a successful L2 access[15].

3) Service Access

With the high number of different authorities operating the WLANs, the homogenization of the provided services seems hopeless. In contrast, the relatively small size of each WLAN makes the frequent network changes quite probable. Moreover, numerous WLANs are likely to co-exist at popular locations. Given such a heterogeneous service environment with competing offers and frequent network changes, it is indispensable to provide an effective user service discovery method.

Existing service discovery protocols include DCDP [110], UPnP [88] and OSGi [89]. Service access issues could also be treated transparently by a middleware (CORBA, Jini, etc. [57]). While the approaches vary a lot, in classical 802.1X architectures they are potentially based on the data service (e.g. transported by the network layer or higher) and thus require a previous station/user authentication. However, in the applied AAA

[15] If the accounting is not activated at L2 access, then the infrastructure usage is free of charge.

architecture the billing of the access is based on the AAA accounting. The latter however starts with the completion of the authentication procedure (as soon as the authenticator opens its controlled port). Thus, the user faces the situation where he has to start paying for something before actually knowing what it is going to offer him (pay-before-choose).

Currently, to be able to use asynchronous services like e.g. Instant Messaging (IM) or SMS-relay, the user would have to proceed as follows:

1. Complete the L2 authentication.
2. Obtain the L3 configuration (e.g. by DHCP).
3. Start the service discovery procedures (could take some seconds).
4. Follow the defined procedure to connect to the service.
5. Send the actual message (SMS or IM message).

That is too complicated given the user's original intention to send a single datagram (e.g. a typical GSM SMS limit is 160 bytes data plus control information as compared to EAP PDU of 1000 bytes).

We have studied IP micromobility integration in such networks [111]. Waking up from the idle mode, the mobile has to redo the whole L2 access procedure. Besides, it is not possible to use the proposed *paging* [112] mechanisms. The paging messages from the network would have difficulties to reach the mobile since the controlled port at the new AP is likely to be in the closed state.

4) Access Point Selection

Due to the relatively small radio cell size of an access point, large scale, enterprise WLANs have to deploy numerous access points to provide WLAN connectivity at a desired location. To provide seamless access, several access points are likely to be available at each position. However, the access point selection is insufficiently covered by the 802.11 standard. The SNR (signal-to-noise ratio) observation which is typically used as the main metric does not always provide satisfactory results [113].

In the AAA-based architecture of 802.1X, the authentication server already has security associations with all access points. It could additionally store their geographical positions. Using authentication and accounting data from the APs, AS is best placed to advise the station to use another available access point for data access. Then, the station could make better decisions relying on both locally measured channel properties and the obtained hints. Obviously, this also requires an appropriate signaling.

5) Other problems

Other potential problems include divers in-session signaling, user QoS negotiation, or user-user signaling (e.g. for network prolongation by peer-to-peer networking as proposed in [62]).

System Architecture

This section describes the proposed usage of the 802.1X reference architecture in the new generalized context. It describes the new roles of the architectural entities and introduces the necessary protocols.

We propose a different view at the 802.1X framework. Though it is currently only used for authentication data transport, EAP actually achieves more: by establishing a channel between the station and the network's management infrastructure it opens a way to

exchange arbitrary data between the user equipment and the provider's control plane. Many a difficulty of the modern communications architecture design lies in the vertical layer interaction. Using EAP, 802.1X provides a layer spanning communication channel. In particular, EAP is IP-independent and can be used with any 802 network (802.1X), but also with xDSL and dial-up links (PPP). It can thus be used with the vast majority of the currently existing user equipment.

We propose to use the EAP channel for generic signaling purposes between the station and the network. Hence, in our approach the EAP-authentication is only one application of the general signaling transport. This is illustrated in Figure V-8. Accessing the wireless network $WLAN_1$ the 802.1X-station TER simply sends its signaling data encapsulated into EAP to the next access point. The encapsulation is done by the means of EAP-SIG, a new EAP method. 802.1X-compliant access points forward EAP-SIG PDUs to the pre-configured AAA server that additionally acts as a Signaling server (SIGS).

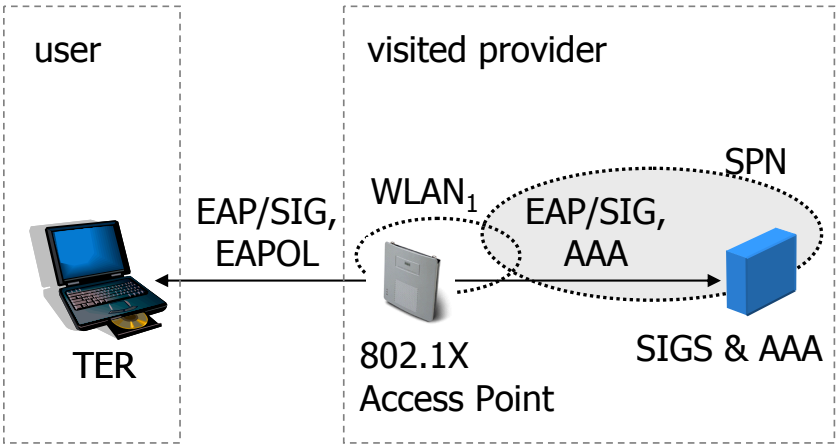


Figure V-8 Main entities and protocols of the proposed solution

Advantages

The proposed extended usage of EAP opens new possibilities providing a channel for signaling exchanges. This approach has the following advantages.

1) Signaling before Authentication

Sending signaling over EAP opens the possibility to exchange signaling data before the start of the actual (complex) authentication procedure requiring a dozen packets and more. This enables e.g. AP selection signaling since the actual AP should be chosen before the authentication with that AP is started. At the same time this enables asynchronous messaging services or general one-frame-signaling (for divers state refresh purposes, like e.g. route update in micromobility approaches, etc). This also allows sending paging requests to the mobile terminals.

Note that pre-authentication signaling does not imply insecure signaling. In our approach, we recommend to encrypt and sign signaling data. This is described in the following sections. If services are used with this mode, the service can be billed directly at the Signaling server.

2) Choose before Pay

Using EAP-signaling the user can request network information before the used access point opens the controlled port and thus before the connection billing starts. Owing to this feature, users can freely browse available networks, searching for a service they are interested in. The Signaling server is able to propose services corresponding to the user's authorization profile. The service description can contain arbitrary information including prices, proposed quality of service levels, etc. Thus, in our approach the user knows what she is going to pay for (*choose before pay*) rather than being forced to pay before being able to select services.

3) Independence of IP

EAP-signaling does not require network access at the IP level. Thus, no IP configuration has to be attributed to the user equipment. This radically limits the user's access to the network. This limiting effect has an important security aspect (since a panoply of attacks is based on IP but no effective attack over EAP has so far been published) but also maintains independence of IP-networking, thus being perfectly suited for other architectures (like e.g. IPX).

4) No periodic messages

Our proposition does not require any signaling-transport specific periodic messages.

5) No pre-configuration

A station accessing a network does not need any pre-configuration in order to be able to exchange signaling with the network's Signaling server.

6) Broadcast-free

In our approach, neither the server's advertisements nor the station's solicitations (both broadcast) are necessary in order to find the Signaling server. The station simply addresses its signaling to the serving access point.

7) Scales with the AAA

For simplification we assume that every network has one Signaling server which is co-located with the Authentication Server. However, neither of these assumptions is required.

Using some protocol between the Signaling server and the Authentication Server, the latter could be remote (e.g. a virtual operator). Such a protocol could be an AAA protocol. Furthermore, for scaling purposes, several Signaling servers could be installed in larger-scale 802.1X networks. The Signaling servers could build a local hierarchy, using e.g. a so-called AAA proxy mode. Hence, our approach scales in the same manner as the AAA-approach.

8) Compatibility and ease of implementation

For interoperability reasons we propose to use the current EAP definition and to transport the signaling data within a new EAP method. Thus, as with the authentication, the access points are not affected by our proposition. The architecture can be added to any existing

network using 802.1X. Our proposition does not require any changes to the EAP carrier protocols (DIAMETER, EAPOL, EAP, etc.)

The popularity of EAP at the IETF proves the relative ease of EAP method definitions. Deployment of new EAP-methods is limited to software installations and is thus simple. It needs an addition of a respective method on the stations and on the AAA-server (support for EAP-SIG).

Technical Details of our Concept

Our approach is based on the introduction of a new EAP method, which we call EAP-SIG. In the 802.1X framework, this method is distinguished from the existing methods by the assignment of a new EAP-Type. This new method has to be supported by all stations, all visited networks and all participating virtual operators. The changes at the virtual operator and in the visited networks are limited to the AAA servers.

EAP-SIG defines two basic modes. The first mode provides a possibility to send an encrypted and signed datagram carrying an arbitrary payload to the visited network. In the second mode, EAP-SIG follows the classical EAP scheme. We call the latter mode the dialog mode. In any case, EAP-SIG acts as an independent transport container for the existing signaling protocols. Typical EAP-SIG payloads could be PDUs of service discovery, micromobility, authentication and various (asynchronous) services. EAP-SIG re-uses the existing security association between the user/station and the network (or the virtual operator) to encrypt its datagrams and protect their integrity. EAP-SIG also supports fragmentation.

The general packet format of EAP-SIG is shown in Figure V-9. The grey fields represent the EAP fields as defined in [94]. The `EAP-Type` field is to be set to the assigned value for EAP-SIG. The following fields include a `Flags` field (1 octet), and a `Subtype` field (2 octets).

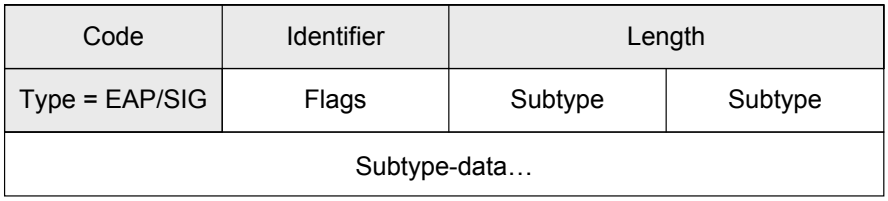


Figure V-9 EAP-SIG packet format in classic mode

The `Flags` field defines the `Fragmentation` flag, which is set on all but on the last packet of a fragmented payload. Since generic signaling protocols are difficult to define, we reserved the remaining seven bits for future development.

Each subtype itself defines an EAP-like payload transporter for one of previously mentioned protocols families. Thus, each subtype includes its own length field. In particular, for authentication purposes, the existing EAP types (e.g. EAP-TLS, EAP-AKA, PEAP, etc.) can be transported "as is". Other protocols are transported starting with `Subtype` values from 256 and above (thus using the first `Subtype` octet). This assures compatibility with standard EAP and allows for a variety of subtypes.

EAP-SIG uses the dialog mode when the pre-authentication signaling is not required. In this mode, EAP-SIG uses its own EAP type and follows the frame format shown in Figure V-9.

For the pre-authentication signaling, EAP-SIG could use its own assigned type. However, that would result in an unnecessary additional round trip of EAP Identity exchanges. We thus use here a further optimization avoiding this unnecessary additional round trip. With that (non-obligatory) optimization, EAP-SIG in the datagram mode does not use its assigned EAP-Type and extends *EAP Response/Identity* (i.e. EAP-Type 1) instead. The latter is classically used to send the user/station identity as a reply to the *EAP Request/Identity* issued by an AP. EAP-SIG adds its payload after the actual identity. The datagram mode does not require fragmentation, so the `Flags` field is zero and thus provides the delimiter. (Note that null-byte termination of the identity string is prohibited by [94] and `Length` field value must be used. This mechanism which we have developed independently is similar to the EAP-based network information mediation introduced in [114]) The resulting datagram is equal to the one shown in Figure V-9 except the EAP-Type is set to 1 and the Identity string is inserted between the `Type` and `Flags` field. The included identity permits the home Signaling Server to verify the signature and to decrypt the appended payload.

A whole network access scenario based on EAP-SIG is shown in Figure V-10. After having established a physical and link layer contexts, user terminal TER uses EAP-SIG in datagram mode to discover network information. It sends its response to the request issued by the access point AP. AP blindly and automatically forwards this message to the local Signaling Server SIGS. SIGS generates a corresponding response and encapsulates it into an *AAA Access-Challenge* message to enable a subsequent TER reaction. That finishes the early signaling phase (pre-authentication signaling). If the received network information is satisfactory, TER can now access the network using a standard 802.1X method as shown in the respective abbreviated phase.

At any given point of time during the session, TER can use EAP-SIG e.g. to discover service properties, etc. SIGS keeps the state of all non-expired authenticated terminals and replies with *AAA Access-Accept* messages for in-session requests.

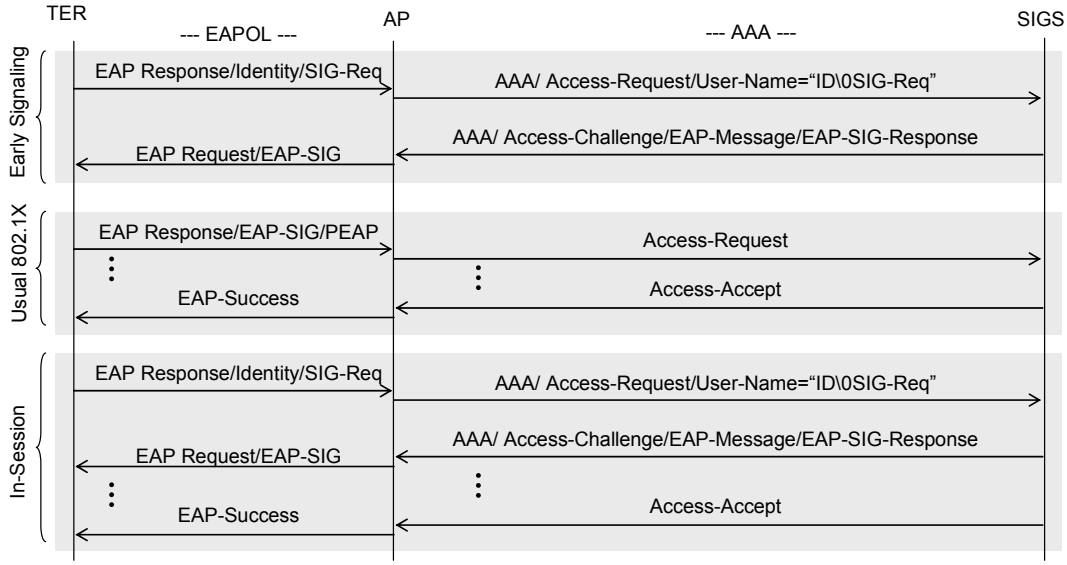


Figure V-10 Successful network access to a network using EAP-SIG

In both modes, the transported EAP-SIG-payload data are encrypted using the existing security association (except for the authentication payload). The encryption is carried out by the encryption algorithm and the encryption key from a security association. If the packet is to be signed, three special subtypes are appended at the end of the packet. The

first (`Nonce`) provides a reply protection. The second (`Visited Network`) displays the name of the used network to the virtual operator. The last (`Signature`) includes the cryptographic signature of the cryptographic hash of the packet content. Signing is recommended in all messages. It is mandatory in the datagram mode. A cryptographic hash (e.g. SHA256) is built over the whole packet starting with the `Code` field, including both `Identity` (if present) and `Signature` subtypes (the latter filled with zeros before hashing). Depending on the available security association, the result is then signed either by using a signature function (e.g. RSA with a private key), or by using an encryption function (e.g. AES) with a secret key over the obtained hash. The result is put in the `Signature` subtype.

Discussion

With this design, EAP-SIG can be used both before and after the L2 authentication. The mode with a classical EAP-authentication previous to EAP-SIG is also supported. In this case, EAP-SIG can use the keys derived during the authentication phase.

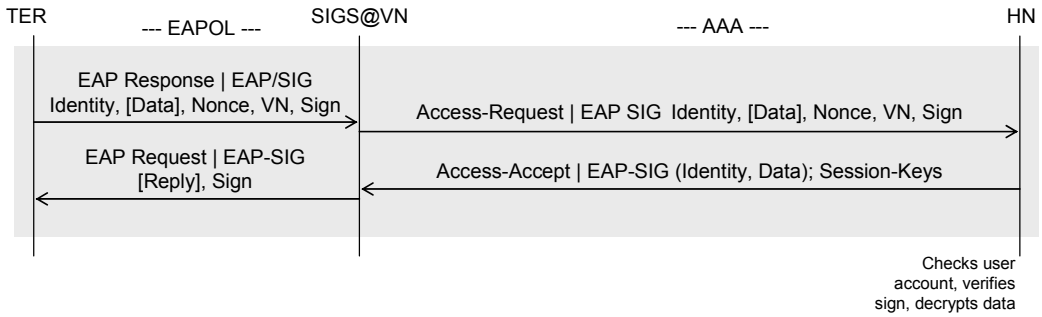


Figure V-11 Successful network access in a multi-domain operation with EAP-SIG in datagram mode

EAP-SIG also provides a datagram service that can be used prior to authentication. The data sent in this mode have to be encrypted and the whole datagram has to be signed. Using the realm part of the provided identity, the visited network's (VN) Signaling Server (SIGS@VN) can forward the packet to user's home network (HN) or virtual operator (see V.1.2). This is shown in Figure V-11. On packet reception, HN verifies the identity, the VN field and the included signature. Using its service agreement with VN it then securely delivers the necessary session keys, authentication tokens and the decrypted message content to the VN. The network operator can thus issue a secure reply to the signaling request using the received session keys. Obviously, this procedure can be repeated several times, if necessary.

Recently, different drafts have appeared at the IETF, using the EAP channel for specific extended signaling purposes. A typical example is the network information mediation pursued by the EAP WG [114], but there are also intentions of standardizing EAP for MobileIPv6 client configuration [115].

From our point of view, these specific usage scenarios are special cases that can – and should – be covered by a generic protocol which can be provided by EAP-SIG.

Security Evaluation

From the security point of view, the dialog mode of EAP-SIG equals a usual EAP operation mode. If a VN does not recognize EAP-SIG, it should follow its policy and act exactly as it acts when it encounters a request for any unknown EAP-type. If the VN

supports EAP-SIG, it must verify that in this mode the first transported payload is an authentication exchange. Since for these purposes EAP-SIG reuses the existing EAP authentication methods "as is", the same requirements apply to the EAP-SIG-transported authentication as for any WLAN authentication (mutual, strong, session key material negotiation, etc. [97]). Note that not all existing EAP protocols are suited for such authentication purposes [81]. The content of subsequent payloads must be encrypted using the negotiated session keys.

The datagram mode raises some new security concerns. Since it allows messages to be sent prior to authentication, an attacker could sniff and modify such messages, create fake replies, impersonate as a network or a user, setup DoS attacks, etc. We mandate encryption of the data in the messages of this mode to prevent sniffing. This also prevents communications with rogue networks. Further, we prevent replay attacks by adding a `Nonce` field. The included signature of the whole datagram enables integrity verification and message authentication. Moreover, since VN can not directly verify the first datagram, the maximum number of datagrams should be limited to prevent DoS attacks.

Figure V-12 shows the same scenario as above with a failed network access. Such failure could be due to one of the classical reasons like unknown identity, invalid account, used-up credit, etc. or because of specific EAP-SIG reasons like wrong signature, failed decryption or invalid encapsulated request, Nonce-replay, VN field not corresponding to the sending SIGS@VN, etc. The exact reason can be expressed by a failure code transported to the VN. Depending on the semantics of the error code, VN can use *EAP Notification* to forward the error code to TER or block out further requests.

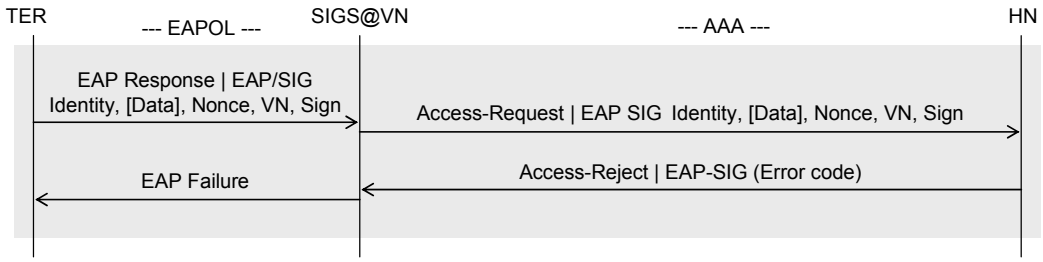


Figure V-12 Home network rejects network access in a multi-domain operation with EAP-SIG in datagram mode

Used in datagram mode, EAP-SIG can not easily provide user location privacy. If user location privacy is required, EAP-SIG relies on temporary identities agreed upon with the virtual operator by other means. Otherwise, a dialog mode authentication is required first.

Prototype Implementation

The prototype implementation is described in detail in Appendix C. Using existing open source, we have implemented prototypes of the user part (EAP-SIG client) and the network part (SIGS) of our concept. These could be successfully tested within the INFRADIO platform (in the Test virtual access network).

Figure V-13 shows a screenshot of the EAP-SIG client after a successful service discovery in a 802.11 network. The discovered services (in the example: SMTP, HTTP, SSH) depend on the used identity. Note that no logical data connection is established in that case. In other words, the 802.1X port on the authenticator – i.e. the access point – is still in the CLOSED state.

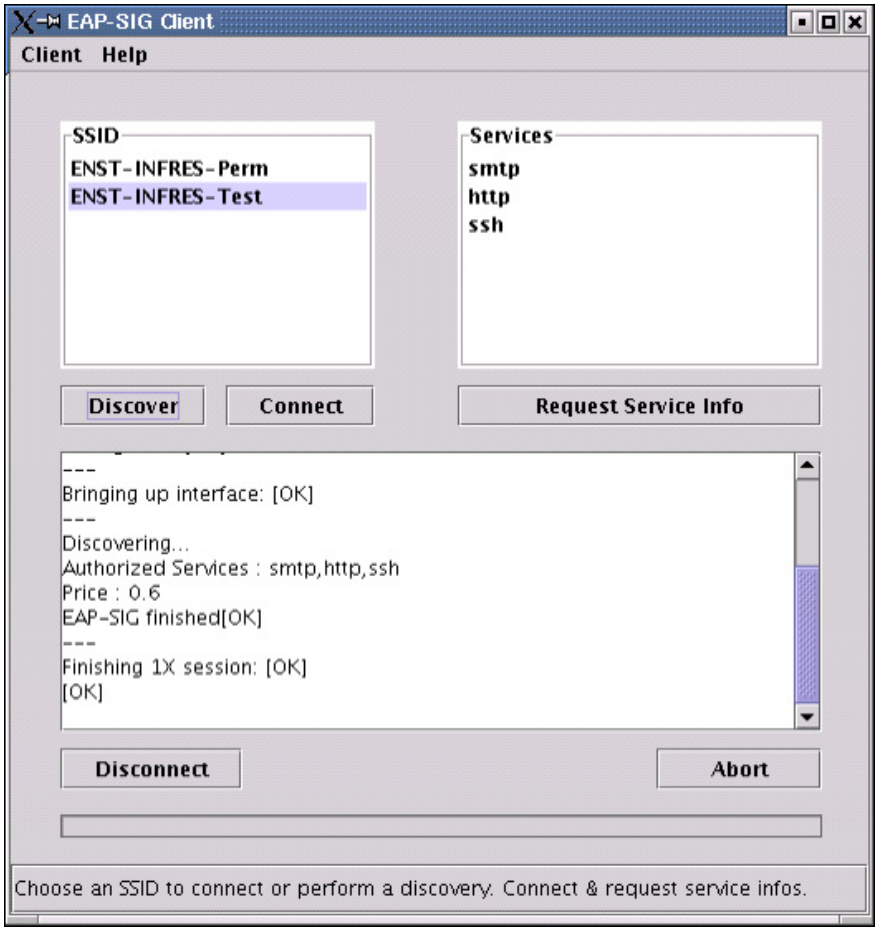


Figure V-13 The developed prototype of the EAP-SIG client

V.5.6. Case study: an Efficient Micromobility Implementation in 802.1X WLANs using EAP-SIG

Motivation

According to Table V-1, the inter-AN mobility within the same SPN could be a critical issue in the 4G networks. While such mobility can be managed by the respective L2 technology in case of a homogeneous handover, with 4G we need to consider new points:

- 4G adds the possibility of a heterogeneous intra-SPN handoff because of the co-existence of different access networks
- 4G potentially increases the probability of such inter-AN handovers (homo- or heterogeneous) by using short-range radios.

Since in our model IP plays an important role, an IP micromobility (see III.4.1) approach seems a suitable candidate to provide a seamless network change. Different IP micromobility concepts have been proposed and studied so far [34][36][37][35]. However, in the presence of obligatory L2 access control, these possibilities are not likely to perform according to the expectations as discussed in the next section with the example of an arbitrary IP micromobility concept and 802.1X access control in a (homogeneous) 802.11 environment.

Then, we show how our generic signaling approach – implemented by EAP-SIG – can be used in conjunction with the mechanisms of the same IP micromobility concept to potentially result in better performance by eliminating latencies.

IP micromobility and 802.1X Access Control

To illustrate the interaction between the IP micromobility solution and the 802.1X network access control, we use an example network presented in Figure V-14. It consists of two IP mobility cells (`cell1` and `cell2`) and an AAA server (`as`) within the local domain. The first cell is managed by the mobility agent `ma1` and contains two access points, `ap1` and `ap2`. The second cell is managed by `ma2` and uses the access point `ap3`. Dashed lines represent AAA connections from the AAA clients to the AAA server; solid lines represent user data links within the cells and arrows show MS' network access. Note that the MAs are typically further connected to some higher-hierarchy MA depending on the micromobility concept used (not shown).

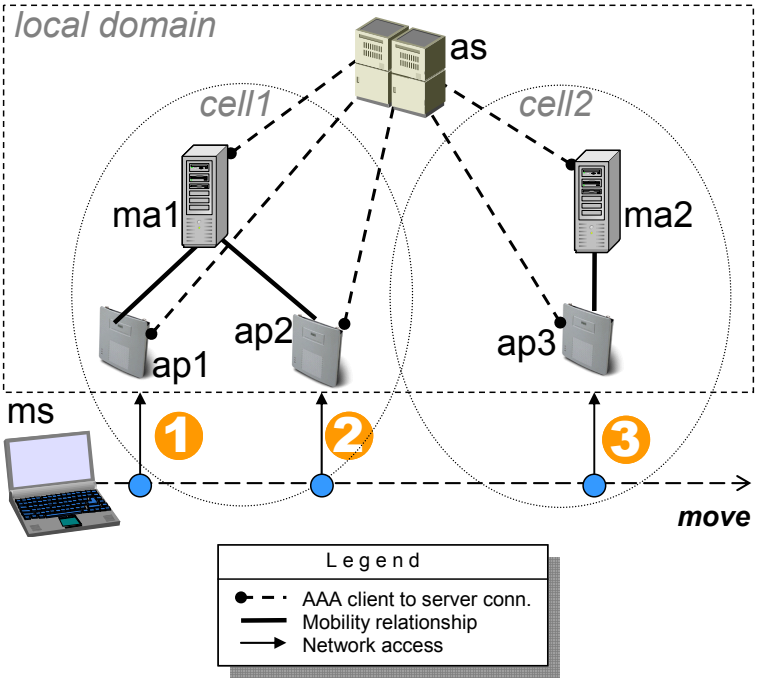


Figure V-14 Movement scenarios in a reference example network

As the mobile station `ms` moves along the illustrated axis, the following three cases occur. The case 1 represents the first network contact, the case 2 represents an intra-cell handover (within `cell1` from `ap1` to `ap2`) and the case 3 represents an inter-cell handover (from `cell1` to `cell2`, i.e. from `ma1` to `ma2`). In the classical integration, the exact sequence of the respective handovers does not substantially change the signaling. Thus, these cases can be used to study every possible scenario with respect to micromobility.

In the specific case 2 (intra-cell 802.11 handover) an integrated L2 context transfer could be used (e.g. IAPP in 802.11 [82]). However, in the general case of a heterogeneous inter-AN handover, such technology-specific integrated mechanisms can not be used. For generality reasons, we thus presume that such mechanisms can not be used. Generally however, the 802.1X delay as such could be minimized. Proposals related to such minimization can be found e.g. in [116][117].

Analyzing these three cases, we show a negative synergy of IP micromobility and 802.1X, i.e. a possible negative effect on both used mechanisms.

Regarding the handover performance, the network access delay is a critical issue. In the L3-mobility supporting networks this delay primarily consists of three phases [35]:

- L2 establishment phase,
- L3 establishment phase,
- L3 mobility phase.

Although all these delays should be kept short, IP micromobility concepts only minimize the third delay. In effect, L2-delay is considered to be near zero [34]. Also, it is generally assumed that the IP configuration must be available before any IP exchanges can occur. Due to the usually necessary DHCP exchanges [107], the IP configuration delay is generally not zero. Here, we do not further consider this delay. Note however, that the IP-layer should not be reconfigured on every handover. Also, it could be easily configured during the L2 access control exchanges (common practice e.g. in PPP [102] – and also possible by using EAP-SIG provided the necessary exchanges). The abstract principle of a mobile network access to a network using 802.1X access control and an IP micromobility suite is shown in Figure V-15. Although the exact exchanges and their interpretation can slightly change depending on the used micromobility concept, the very principle of the illustrated phases remains valid. Most notably, the respective phase latencies always sum up to a common total (parallelization is not possible because of the strictly "layered" system design). We detail the main problems for every case shown in Figure V-14.

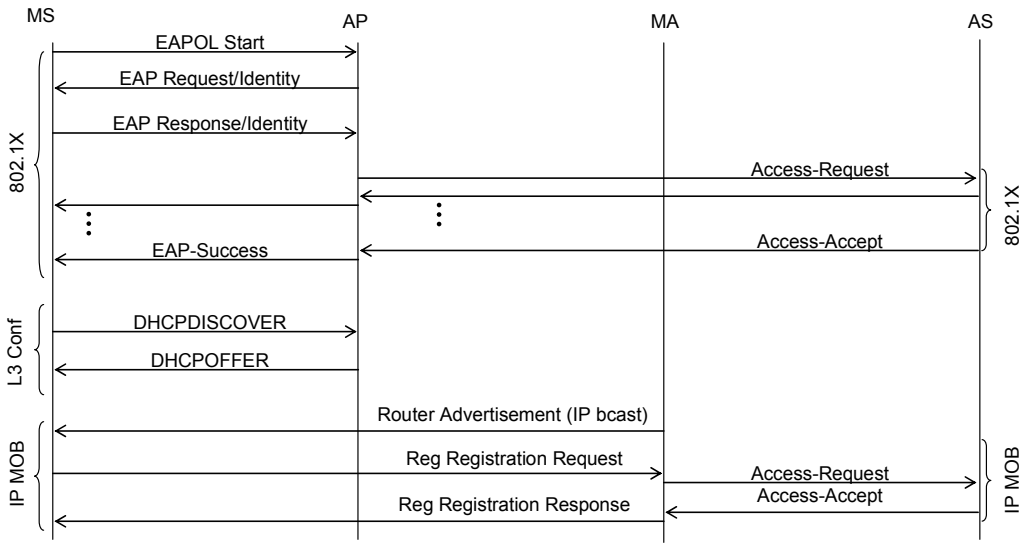


Figure V-15 Schematic flow-chart of the network access to a WLAN classically integrating 802.1X access control and IP micromobility

We generally argue that the L2-delay – otherwise considered insignificant [34] – can become critical with 802.1X. Analyzing case 1, we notice the following two points. First, relying on the AAA-data supplied by `ap1`, the AS (or, more abstractly, the network) can conclude which user is connecting over which AP (see Figure V-16: the `User-Name` attribute contains the user identity, the `Called-Station-Id` attribute contains the unique MAC address of the used access point, the `Calling-Station-Id` attribute contains the unique MAC address of the user's terminal, etc). Thus, directly after the 802.1X exchanges, the network already has the IP location data required for

micromobility routing. Yet, MS and MA start an unnecessary additional mobility registration process because the mobile agent `ma1` is not informed of the current position of `ms`. Secondly, if the mobility solution itself applies an AAA-based access control, `ma1` contacts `as` for the second time for the same terminal to network access.

0	8	16	31
Code: Access-Request	***ID:*** 42	***Length:*** 121	
Authenticator: 0xe16c8f1a3d9326a9025fb043c7f2ecec			
***List of attribute-value pairs**:* *User-Name* = “userA“ *NAS-IP-Address* = 10.10.10.1 *Called-Station-Id* = "00:40:96:35:be:d6“ *Calling-Station-Id* = "00:40:96:42:6f:05" *NAS-Identifier* = "nas1" *EAP-Message* = "\002\000\000\n\001userA"			

Figure V-16 Typical RADIUS packets contain user to location mapping data

Now, consider case 2. While changing from `ap1` to `ap2`, `ms` generally encounters an unauthorized port at `ap2` and thus proceeds in exactly the same manner as in case 1. The 802.1X-authentication between MS and AS has to be repeated completely (or partially, when the used EAP method provides for faster consecutive authentication, so-called *session resumption*). Furthermore, due to the effectively occurred L2 reconnection, MS sends a solicitation message or waits for an advertisement of an MA, even though the cell has not been changed and thus no location update is required. Although the overall delay is roughly the same as for case 1 (except for slight changes in the mobility signaling), case 2 is logically even less optimal: whereas in case 1 some procedures are indispensable (like the first authentication), in case 2 these are either unnecessary (e.g. micromobility-related signaling) or could be significantly shortened.

In case 3, the exchanges essentially remain the same as in cases 1 and 2. Due to the effective cell change, mobility signaling is appropriate. The 802.1X authentication delay should be reduced by using an optimized EAP method providing for session resumption (exactly as in case 2). Similarly, when AAA is used by the MAs, the load increases at AS (as in case 1). Alternatively, the MAs could communicate with each other (or with the higher hierarchy element) in order to verify the presented credentials. However, this mode is not supported by every IP micromobility concept.

Existing EAP-methods providing strong mutual authentication and session keys (EAP-TLS, EAP-TTLS, etc.) typically require a dozen of packets between MS and AS. Even with session resumption, EAP-TLS still requires six packets (EAP-TTLS, PEAP: 8). TLS-based EAP-methods especially can potentially result in more packets due to the fragmentation when using big X.509-certificates. Furthermore, `ms` could be a mobile host from a different administrative domain (roaming user). In such a case, AS can not verify the presented credentials locally and has to contact a remote AAA-server first, thus significantly increasing the 802.1X-delay. With roaming and the related remote authentication service resulting in round trip times (RTT) of ~100-200ms, L2-delays can go well beyond 1s on every handover. It is unclear if the micromobility solution will still be able to counter such a drastic increase.

According to [118], the AAA infrastructure must not be contacted on every handover since it is not meant to be used in such an aggressive way. However, using a standard 802.1X architecture, this requirement is violated in both intra- and inter-cellular handover. Given the potentially high handover frequency with micromobility and the fact that the most popular AAA protocol (RADIUS) is known to exhibit a degraded performance under a high load [91], the AAA infrastructure could get overloaded thus temporarily disabling any network access.

The proposed optimizations [116][117] can help to eliminate the (unnecessary) re-authentication against the AAA infrastructure in case of a local handover. However, the AAA contact could also be provoked by the micromobility concept. Besides, these optimizations are tailored to a specific L2 which does not correspond to our general intentions. Also, they typically require equipment changes and are not yet available.

We observe that the overall handover delay is increased in all cases due to the 802.1X (re-)authentication, double AAA-usage and unnecessary L3 exchanges. Quantitatively, this can differ depending on the following criteria:

- used EAP method,
- round trip time from APs and MAs to the AS,
- IP micromobility solution and its integration with AAA,
- average AS load.

Recent publications show a disastrous performance for standard Mobile IP handovers in a real practical 802.1X environment [119][120]. However, it must be noted that these do not use the proposed optimizations to the 802.1X handover.

Being primarily oriented towards its main purpose (access control for 802.1X, fast handover for IP micromobility), both 802.1X and IP micromobility are aggressive in terms of signaling and repeating exchanges. We conclude that a direct integration of 802.1X and micromobility can have a negative mutual effect on the performances of both used mechanisms. The core problem is an independent user location tracking in both mechanisms. In the next section, we show how EAP-SIG can be used to overcome some of the drawbacks illustrated above.

Proposed Solution

We propose a different approach to micromobility integration in the networks applying 802.1X access control. Our proposition aims to reduce the L2-delay that becomes significant with 802.1X. At the same time, it conforms to the requirements defined in this paper.

Our proposition is based on two observations. First, all standard 802.1X APs blindly forward every packet received over the uncontrolled port to the preconfigured AAA-server. This permits reaching the network's control entity without any additional, delay-prone L3 service discovery messages (router solicitation, advertisement, etc). Second, the 802.1X access control implicitly retrieves the data necessary for micromobility routing (see Figure V-16). These data could be used directly by the micromobility agents. In order to exploit these properties, we propose a slightly different architecture using EAP-SIG as micromobility protocol transporter.

Operation Mode

In our architecture shown in Figure V-17, each original MA is replaced by a micromobility-enhanced AAA proxy-server (MAAAP) implementing MA's standard functionality and an EAP server with AAA proxying support. Accordingly, the AAA

clients of the APs are configured with the addresses of their respective cell-MAAAP (i.e. `ap1` and `ap2` with `maaap1` and `ap3` with `maaap2`). Instead of separated 802.1X and micromobility clients, our MSs use an EAPSIG based client performing an integrated network access and mobility control. Note that, conforming to our requirements in [111], no modifications are required on the APs since EAP-exchanges are transparently resent to the MAAAPs. Following our configuration, the APs are organized in logical groups corresponding to their cell assignment.

EAP-SIG transports different signaling payloads. Their exact content depends e.g. on the used authentication and micromobility concepts.

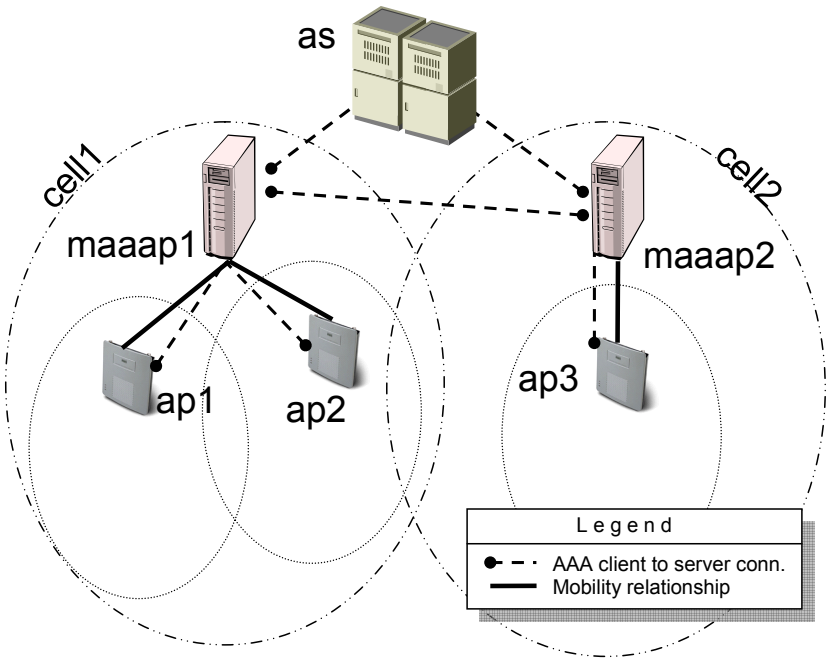


Figure V-17 Proposed integration of IP micromobility in 802.1X WLANs

If necessary, MAAAPs can translate the incoming EAP-SIG-messages in the respective protocols and forward the resulting messages to the appropriate entities (e.g. to the `as` or to other MAAAPs or even standard MAs). We use EAP-SIG as a L2-transport mechanism for the required signaling, in the same manner as EAP is used as a transport for arbitrary authentication payloads. We additionally define the following requirements for the payloads transported by EAP-SIG in this specific scenario:

- Transport of authentication payload: the authentication method should support strong mutual authentication, session key (SK) exchange and SK-based, fast re-authentication (*session resumption*) for both L2-delay and AAA-load minimization.
- Transport of the additional micromobility-relevant data from MS to MA: address of the previously used MA, current MA, handover-related signaling.

An example flow-chart of EAP-SIG micromobility exchanges in a scenario with Cellular IP (CIP [34]) is shown in Figure V-18. MS uses EAP-SIG to transport TLS for user authentication. EAP-SIG permits micromobility signaling to MAAAP and authentication with AS. MAAAP extracts and forwards TLS frames to AS using the AAA protocol. The latter finally informs MA of the successful user authentication (`Access-Accept`)

providing it with a temporary MAAAP-MS security association (`Session SA`) that is to be used for a local re-authentication.

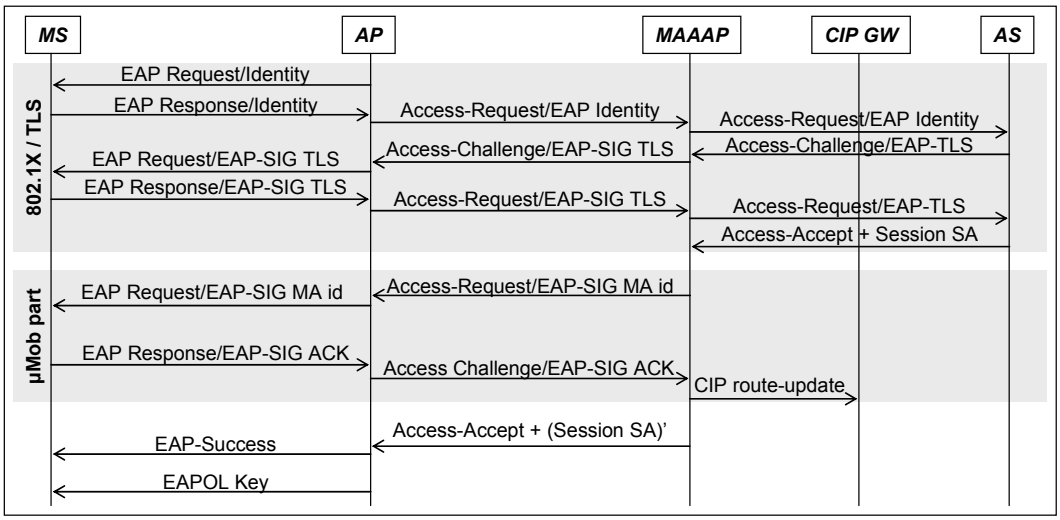


Figure V-18 Mobile network access with EAP-SIG (example with Cellular IP)

Further, MS is informed about the MAAAP's identity. MS saves the current MAAAP identity and uses it in the subsequent handover signaling. Having received an acknowledgement, MAAAP sends an `Access-Accept` message to the AP adding a temporary AP-MS security association (`(Session SA)'`) derived from the MAAAP-MS session SA. This is actually normal behavior for any AAA proxying server (see e.g. V.4). At the same time, MAAAP can send IP mobility signaling to other mobility agents. In Figure V-18, MAAAP starts CIP [34] signaling to the CIP gateway (CIPGW) in order to change the per-host routing for the new user.

Note that the two added micromobility-related messages ("µMob part" in Figure V-18) hardly produce any additional delay since they are exchanged with a dedicated mobility agent in the local subnet (MAAAP). Furthermore, this additional exchange could be used to transport supplementary data, e.g. IP configuration, security associations, etc.

Analysis of our Proposal

In this section, we show how our proposal outperforms the classical integration in all three cases.

First, we analyze the case 1. As MS connects to `ap1`, it is requested to start the L2 authentication signaling. Following its new configuration, `ap1` relays the presented identity data to `maaap1`. Since `maaap1` generally can not verify the credentials locally, it relays the packets between the network's AAA server (`as`) and the MS. Following [118], the AAA-infrastructure should be responsible for assigning an IP address to the new connected MS. Alternatively, each MAAAP could be assigned an address pool from which it could assign addresses to MSs. Necessary mechanisms exist e.g. in RADIUS. Regardless, the current MA (`maaap1`) knows which user/IP is connected over which AP not later than with the final response of AS. It can thus start the necessary micromobility signaling with the micromobility supporting network using one of the micromobility protocols ([34][35][36][37]). According to the specified methods, MAAAP performs local operations (e.g. updating the user's location and the routing tables, establishing new tunnels, etc.) In some cases, MAAAP can start this signaling even before the AP opens

the port, i.e. before MS has L3-access. Thus, these two latencies do not simply sum up as described above for the classical "layered" integration. Moreover, compared to the latter, no additional (post L2) micromobility signaling is necessary on the MS-MAAAP interface in the access phase (like e.g. broadcast-based router discovery signaling, etc.). Thus, the L3 mobility delay is decreased or even eliminated if the micromobility procedures can be accomplished prior to L2-authentication end. The AAA infrastructure is not (unnecessarily) exposed to an additional load. The overall handover time can be minimized roughly to the standard L2 handover delay.

In case 2, `ap2` first requests MS's identity. Using session secrets and fast re-authentication supported by EAP-SIG, MAAAP and MS can perform a fast local authentication without repeating the AAA procedure. Except for the fast handover support presented later, the L3-micromobility signaling on the MAAAP-MS interface and the L3 mobility delay can be completely avoided in this case.

In case 3, EAP-SIG is used to transport the identity of the previously used MA to the new MA. As a result, `maaap2` contacts `maaap1` for identity verification. This avoids an additional AAA contact and thus fulfills the requirements in [118]. This also further minimizes the handover latency since generally `maaap2` is likely to be closer to `maaap1` than the central AAA server.

Hence, in all three cases the usage of our architecture results in a decrease of both handover-latency and AAA-load as compared to the classical integration.

Handover Optimizations

Fast and smooth handovers (HO) are designed to optimize the HO latency and the HO packet loss respectively. We show two ways how fast and smooth handover can be implemented in the described environment.

Every active 802.11 AP can emit so-called beacons [14] containing the L2 address (MAC) of the heard AP. An MS receiving a stronger beacon from a new AP can send a notification of an imminent handover to its current AP by using EAP-SIG. This message essentially contains the L2 address (MAC address) of the new heard AP. The current AP automatically delivers this message to the current MAAAP that is capable of assigning the included L2 address to the respective IP mobility cell. From here on, full optimized handover support of the respective micromobility solution can be used in order to notify the new MA/MAAAP about the imminent handover and to activate packet stream diverts, buffering, etc. – as described by the micromobility solution.

An alternative scheme can also be used with EAP-SIG. MS sends an EAP-SIG packet containing the old MA's identity through the new AP to the new MAAAP. The latter verifies the contained user identity by interrogating the old MAAAP using micromobility signaling and authorizes the new AP to open the controlled port. Such reactive handover is generally inefficient for purely IP-based methods since a delay-prone L2 access has to be completed first. Since the EAP-SIG traffic is exempted from the port control and no agent/router discovery is required, the new MAAAP can be immediately informed.

Conclusion

In this case study, we address the integration of IP micromobility in the access networks using L2 access control. We show that a classical "layered" integration can result in a degraded performance of both mechanisms in terms of latencies and the associated packet loss. We present a micromobility- and hardware-independent, easy-to-implement integration scheme which recovers the full performance of both mechanisms.

The key innovation of our proposal consists of the generalized usage of EAP for signaling purposes (EAP-SIG, see V.5.5) and the introduction of the new architectural element (MAAAP) translating network access data to a micromobility protocol. Our scheme enables an easy micromobility upgrade of a deployed network by installing conforming clients, mobility agents and slightly adapting the AP configuration. We address such specific micromobility-related issues as seamless handover and fast security support and show how these can be implemented in our architecture.

Due to the generality of the 802.1X concept and its similarity to the dial-up model, this solution can be directly applied to a variety of popular network technologies.

V.6. Conclusion

In this section, we discussed the new possibilities to provide lower layer signaling on the user link in the 4G scope. We conclude that the higher level methods are not efficient while the lower layer mechanisms are technology-dependent and thus lack integration.

We thus proposed a new system access model featuring two distinct channels. With this abstract channel definition, we are able to push the user signaling away from the access control equipment towards the control plane of the used SPN. That is what we call the virtualization of the signaling. Instead of defining what is done, we only define how it is done, leaving the implementation to the respective application layer. In this model, the user signaling remains independent and transparent from the access controllers and even from visited networks (that can also act as blind pass-through to the home network). The protocols necessary for the access device to the SPN control plane communications are network internal and thus can be proprietary. Our approach thus loosens the constraints where these are particularly tight: in the weak user terminals.

To realize our novel ideas, we propose a generalized interpretation of the data exchanged over EAP. We introduce EAP-SIG, an EAP-based signaling transport mechanism for future telecommunications architectures. Our approach is extensible as it is capable of transporting any signaling payload. Introducing payload encryption and signing, it integrates security measures.

We implemented this mechanism in our test platform. Implementation details can be found in Appendix C. Our approach does not require any changes to 802.11 and 802.1X standards. Principally, EAP-SIG can also be used for user signaling purposes over all dial-up networks.

As a case study we analyze the inter-AN mobility support in our 4G vision. We show how EAP-SIG can help to reduce latencies when integrating IP micromobility in access networks protected by L2 access control mechanisms.

CHAPTER VI

Future Network Architectures

Different users need different services with different guarantees. Depending on the services and guarantees, different access technologies must be used. The choice of the access network technology must be met within the user terminal (with user assistance or by a pre-configuration) by investigating its communicational environment over all available physical interfaces. However, the access technology and its capabilities is not the only decision criterion. If different access networks of the same technology are available, other criteria must be used to build the decision base. In essence, this is the user view of the *network selection problem* (see Section IV.4.1).

The same issue can also be addressed from the provider's point of view. If different users need very different services, then providers have to deploy different infrastructures. Clearly, each provider is interested in using the available infrastructures so that as many different service scenarios as possible can be supported by the latter. That allows cost savings, new business cases through increased flexibility and guarantees revenue.

That is why a major part of this work is dedicated to the definition, development and analysis of the architectures and mechanisms that can provide for the maximum possible infrastructure reuse. Such adaptation of ANs and SPNs is an abstraction of the real physical infrastructure. The network functions and services are virtualized in the sense that these are now defined as interfaces, in a way comparable to the virtual function definitions in high level programming languages such as C++ or Java. This decoupling of the implementation and the interfaces is what we call *virtualization* of infrastructures and

services. In our user-centric 4G-vision, the user profile and the criteria from the provider policy (including e.g. such issues as the current network load, etc.) are the only input structures to do that.

In this chapter, we show how such virtualizations can be practically achieved at different OSI layers (see Section III.2). In the following section we treat the virtualization of the physical access networks using the example of the deployed and fully functional wireless environment. In the subsequent sections we treat two approaches to the virtualization of the services' infrastructure within a SPN showing our ideas examinded and implemented in two research projects. Finally, we address the organization of the SPN's control plane motivating alternative approaches to the currently predominant centralized concepts. Our solution is an adaptable, scalable and self-organizing control plane architecture.

VI.1. Virtualization of the Physical Infrastructure

VI.1.1. Introduction

In this section, we study the current possibilities of the infrastructure virtualization. As an example we use a physical 802.11 WLAN in which the access points are interconnected by a Ethernet-based (100BaseT) switched physical network.

In the RNRT project INFRADIO [30], we had the opportunity to study our strategies and our general approach to network deployment. The first phase of the project included the deployment of a fully operational and extensible wireless LAN integrating services. We used the experiences gained in the AUTHENTIS project (see Section V.4).

The results of this work are detailed in [121] and [122]. In the following sections, we present the encountered network environment, our security policy, the identified issues with the usage of the access networks within the environment and the solution applied in the operational network.

VI.1.2. Network Environment

A university campus is a highly open network environment with typically loose to non-existing security policies. Each department has its own network which is very tightly coupled to the university backbone. Decisions on who has to authorize which action are thus often unclear. Besides, different network and service technologies are used in parallel. New test networks are often simply plugged into the physical infrastructure without any authorization. With wireless technology, this potentially enables anyone to access the network core services.

Another interesting problem to study in that environment is the deployment of the access network. From our point of view, above all it includes the study of the integration possibilities of the access network in the SPN (data plane connection to the SPN, control plane connection, etc.). Also, the internal trust relationships between the AN and the SPN in general, and between the controllers and controlled entities have to be clarified, justified and defined.

VI.1.3. User Groups

The user groups vary from permanent university personnel of different levels (professors and researchers along with the administrative personnel) over long-term guest researchers and PhD-students to graduate students and, finally, daily visitors or guests. Thus, the potential user groups are very different in terms of attributed trust and service access authorizations. The user groups also differ in terms of the associated commitment of the network administration: while the services must be delivered securely to some users (e.g. professors, personnel, researchers), guest services do not imply any guarantees.

Additionally to system administrators, in our analysis we have defined two major system user groups which have very different trust levels. These groups are:

- Users, further subdivided in permanent and temporary users,
- Guests, further subdivided in visitors and real guests (see Section V.1.2).

In the user group, we further distinguish different subgroups according to their service access needs. We consider the access to different network internal services (Intranet) and the Internet-service.

VI.1.4. Security Policy

In the preliminary complete security analysis, we have compared the strength of different available integrated WLAN security mechanisms, including different EAP authentication methods, link encryption and frame integrity mechanisms. We have also studied the possibilities of the VPN technology as a possible candidate. We have defined the security objectives and the corresponding subjects and objects. The security architecture is part of the network architecture; both have been designed simultaneously.

To achieve a reasonable overall security level in the encountered low-discipline environment, we pursued a twofold approach. On one hand, we have worked out a user service agreement document containing the basic user rules. This document is acknowledged by the permanent users. It is also acknowledged by the guests and visitors to whom we display it on every connection. It gives us the necessary authorization but most importantly also informs users about their possibilities. On the other hand, we pursue the user seduction strategy trying to attract all users to our network.

Consistent to the overall university security policy, the permanent users are fully trusted in our architecture. They have the same unconstrained connectivity as in any wired network of their respective department. However, due to the medium change, we had to change the security policy in terms of user identification. Access as a permanent user has to be subject to a strong L2 access control (see Section IV.4.1). Various L2 authentication methods should be supported for different subgroups of the permanent user group. Visitors and guests should be granted free but also very limited access. A user-oriented authentication procedure should be supported for an extended guest/visitor service[16].

VI.1.5. Problem Statement

The different assigned trust levels result in different user authorization sets. In particular, this applies to the deployed infrastructure. Simultaneously, the physical infrastructure protection doctrine in the wireless networks (see Section IV.4.1) requires L2 access control.

[16] For practical reasons, it is important to provide guests an access method which can be configured and activated without requiring local administrator rights on the user equipment

The above two points require a consistent user diversification from the beginning, including the L2 access control. Hence, the AN access control measures have to support different methods for the treatment of different users and their respective data flows. For instance, the provided access should differ in terms of obtained connectivity, network neighbors, connection quality (e.g. in terms of bandwidth), etc. That represents a technological problem.

VI.1.6. Network Architecture

The virtualization approach in fixed networks can be implemented using tagging or labeling mechanisms. Generic mechanisms exist and are widely used by major telecommunications providers as well as in the local networks. MPLS [73] is a good example for a mechanism for building virtual infrastructures (e.g. links or connections) over physical networks. In local Ethernet networks, VLAN tagging (IEEE 802.1q [123]) is usually used. VLANs can be seen as a special case of the more general MPLS approach. This mechanism is based on a special tag field (VLAN identifier, VID) that can be added to the Ethernet MAC frames. The switch configuration and the included tags change the forwarding behavior of the latter creating virtual paths through the same physical mesh. 802.11 access points can map the logical ESSID membership [14] to a chosen VLAN tag on the wired links, thus acting as 802.1q-capable bridges.

Each physical AP thus provides different connectivity services in form of diversified access possibilities to various VLANs (see e.g. [121] for more details). This technique is often referred to as *virtual AP*. Using the latter we provide users with multiple parallel wireless networks with different security and link-QoS properties over the same physical infrastructure. This is represented in Figure VI-1.

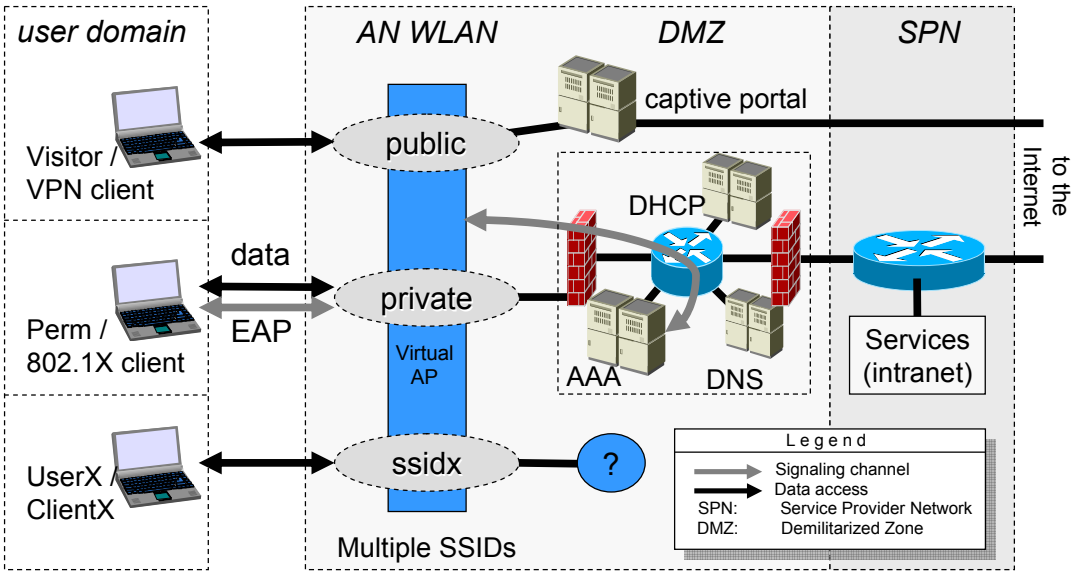


Figure VI-1 The network architecture of our network

The deployed physical AN replies to the requests for different networks, thus providing multiple virtual wireless networks (using the ESSIDs public, private, etc). The APs tag the packets respectively separating them in multiple virtual wired networks using IEEE 802.1q tagging. One additional VLAN (management VLAN) corresponds to the AP's own traffic; it carries control plane traffic exclusively. In our configuration, all virtual APs build one (virtual) L2 segment, i.e. only L2 handovers occur when users change the

point of attachment. We treat each L2 segment differently, both in terms of routing/switching and L2 properties. Each virtual local area network is described in detail in the following section.

Management and Private VLANs

Management VLANs transports control plane traffic to and from the APs. The private VLAN is the virtual network for the permanent users. Both VLANs are switched to our demilitarized zone (DMZ). The latter strictly separates these ANs from the SPN. It is protected by two firewalls and contains user services like Web proxy and SMTP relay, core network services like DHCP and DNS, control plane elements like AAA server, public key infrastructure (request, certification, repository), network observation and management tools (intrusion detection, time servers, logging facilities), TFTP server for configuration/firmware storage. It also holds a part of the business plane, notably tools such as administration console and user and AP management databases. The management VLAN provides the connectivity between the physical APs and the control entities in the DMZ. The wireless part of the private VLAN is subject to a strong L2 access control. With our experiences (see V.4), we use IEEE 802.1X access control with different authentication methods for that purpose. Table VI-1 summarizes the method per user category. Besides, IP packet filters limit the users of the private VLAN to the network core services (DHCP/DNS) and the user services only, merely routing their traffic through to the SPN.

Our system allows for the different digital trust representations to be used by users within the group of permanent users. According to our definitions introduced in Section V.1, user identities in our system are pure system identities. Thus, independently of the used trust representation (certificates or shared secrets) we always separate the system identity (user) from the legal person, machine, etc. behind it. This is automatically given with shared secrets. However, the certificates tend to contain a lot of additional information. Reflecting our definitions, our certification policy is very light-weight. This light certification policy uses standard X.509 certificates for compatibility but is strictly limited to the verification of the user identity correctness, without any assumptions about the related authorizations (as e.g. account validity). As in the case of shared secrets, we try to store all authorizations in the user profiles, within the User Database. This policy allows us – to a big extent – to avoid a lot of classical PKI problems e.g. certificate revocation, re-certification in case of authorization change, etc [100].

Table VI-1 User groups with the implied access control and user authentication methods

User groups		Access control	Transport	Auth method
Permanent	Personnel	802.1X	EAP	EAP-TLS, PEAP
	Guest researchers, PhD students	802.1X	EAP	EAP-TLS, PEAP
	Master, graduate students	802.1X	EAP	PEAP
Guests/ visitors	Without login	L3 forced redirect	-	-
	With login	L3 forced redirect	HTTP	UAM
	VPN users	VPN	PPTP	MSCHAP

Public VLAN

This VLAN implements a hotspot sub-architecture for the deployed network. It is supposed to be used by guests and visitors. We switch this VLAN to an isolated router called *captive portal*. This implements core network services and all higher level guest services. As the first router for the public VLAN, it redirects all incoming HTTP requests

to a web server (co-located) which provides public site information for guests and visitors. A login page is available from there. This mainly implements the Universal Access Method (UAM) as recommended by the WISPr recommendations [124]. We additionally provide a PPTP-based VPN service [125] that can be optionally chosen from the login page.

The login page gives different access levels (see Table VI-1). If the login is skipped, the access is restricted to the public web sites of our platform (containing practical user information on possible services, user manual, troubleshooting and emergency contacts, etc.) and of the ENST (public university web site). If the login credentials are given, a limited Internet access (HTTP/HTTPS, outgoing PPTP and SSH) is possible over a router that performs per-session IP-filtering, network address translation (NAT) and content control through a transparent content filtering Web cache. During the login, users can additionally activate a VPN service for their session. This VPN terminates at the router and primarily serves to protect the user data on the wireless link. However, outbound VPN connections over the router are also permitted, thus providing visitors a roaming access to their home networks.

It is important to notice that normal guests/visitors do not have a valid login because of missing trust relationship. Without this login, the guests would be limited to access the internal pages. To discourage attacks, we want to provide a better service to visitors. However, in the open environment, the trust establishment is difficult. We thus use the concept of *delegation*. We define a special user authorization that enables the holder to create guest accounts without having administrative rights. This authorization is given to some permanent users, e.g. to professors. The created accounts have an expiration clause. Per default they automatically expire at the end of the day. Additionally, our system supports automatic bulk account creation for the invited groups (e.g. an international conference organized in the facilities of the ENST).

The host is often the only person who has a real trust relationship to the invited guests. The delegation concept uses this trust and the trust to the permanent users to give guests an extended system access right. This concept is well suited to relieve the system administration of the frequent short-term administrative burden distributing the effort to the responsible persons.

Other VLANs

Multiple other VLANs are in use for testing purposes, used by different users. Currently, a test VLAN connects to a separate test platform for IPv6 services with MIPv6 support. Our test EAP-SIG implementation (see Section V.5.5) is used in the same VLAN.

The SPN in our model corresponds roughly to the ENST's core network with all contained services, like e.g. the Intranet. A 100 Mbps Ethernet link connects our DMZ to the SPN. The captive portal has a specially treated, dedicated Internet connection.

Major VLANs with the corresponding parameters are summarized in Table VI-2.

VI.1.7. Flexibility of This Approach

The deployed network is designed to support an easy addition of new wireless networks. This is not considered as a daily administrative task. However, given the encountered open campus environment (see above) and the seduction strategy defined in the security policy, an easy and secure addition of new wireless networks should be possible to reply to the extended user needs. Indeed, the research teams are interested in testing new services and ideas. A deployment of an unwanted parallel physical network is less

efficient (because of a probable radio interference). What is more important, such deployment is almost guaranteed to be insecure: in a typical test environment, the first things to be treated lightly are the security aspects. It is also important to provide a secure playground for the testing of new security approaches as such. Finally, new and still very academic quarantine environments [126][127] also rely on a separation of the quarantined entities by some means. We can provide such a separation from the MAC layer on, in an assisted deployment within our managed platform. Such assisted deployment has been performed within the same physical infrastructure merely by reconfiguring several times since the start of the project. An example is the Test VLAN.

Table VI-2 Parameters of the VLANs deployed in our platform

Properties		Defined VLANs			
		Private	**Mgmt**	**Public**	**Test**
SSID mapped?		Yes	No	Yes	Yes
L2 properties	SSID	ENST-INFRES-Perm	-	ENST-INFRES-Invit	ENST-INFRES-Test
	Access control	802.1X/EAP	-	No	802.1X/EAP
	Auth methods	TLS PEAP/MSCHAPv2	-	802.11 open	MD5
	Auth server	DMZ Auth Server	-	-	Test Auth Server
	Encryption	WEP128, TKIP	-	Off	Off
	Integrity	With TKIP	-	No	No
L3 properties	Attachment	DMZ	DMZ	captive portal	Isolated test router
	NAT	No	No	n:1 SNAT (masq)[1]	No
	Access control	IP filter	IP filter	forced IP redirect UAM restricted IP filter	No
	Air link VPN Auth/cipher	No	No	PPTP (optional) MSCHAP/ MPPE128	IPv6 IPSec possible
Services	DMZ	User services, DHCP, DNS	Yes	No (phys. sep.)	No (phys. sep.)
	Intranet	Yes	No (filter)	No (phys. sep.)	No
	Full SPN	Yes	No (filter)	No (phys. sep.)	Yes (IPv6)
	Internet access	Full	Vendor sites (HTTP)	With login: HTTP, HTTPS, SSH (Web cache, content-filter) Else: internal Web	No

The deployment of a new architecture in this physical infrastructure implies several major administrative steps like configuration of a new 802.11 SSID, mapping it to a VLAN ID, VLAN ID switching configuration, a new traffic sink installation, etc.

Typically, the necessary modifications (excluding the installation and configuration of the traffic sink) require less than half an hour of an administrator's work.

VI.1.8. Results

The practical daily experiences with the deployed architecture show so far satisfactory results in terms of usage and security objectives. The platform is fully operational since

[1] *n:k Source Network Address Translation* (SNAT) is the IP address substitution in which a SNAT capable router substitutes *n* source address(es) by *k* other source addresses. *n:1* SNAT is often referred to as *masquerade* because the SNAT router hides *n* different hosts behind one IP address.

summer 2004. The untested elements will be soon transferred to the already existing test environment (currently used as IPv6 and IPv6 mobility test platform).

The developed platform is highly flexible enabling support for very different user groups. These are supported by the virtualization of the provided service from the MAC layer on. Therefore, we are capable of providing links with different parameters depending on the user access level.

However, the network architecture also enables support for very different research tasks. By treating the WLAN as an access network, we are capable of compiling precise user mobility traces in a campus WLAN. Besides, we could easily gather (impersonal) statistics on network usage and verify the results published so far [33]. The possibility to deploy new test environments is currently used by two different research teams. Both environments are virtual and isolated from the rest of the network, interconnecting only the participating external entities through the virtual paths in our infrastructure.

In the immediate future, we plan to extend our architecture to the whole university. The integration of new access points in our environment is easy, since the VLAN technique is already in use in the switches of the core ENST network.

VI.2. Virtualization of Networks and Services

In this section, we concentrate on the higher levels of the OSI stack applying the same virtualization principle to the SPN environment.

The proposed SPN architecture consists of several entities acting in different planes and communicating by the means of protocols depending on the plane and the nature of the entities. Our general architecture distinguishes three planes: a business plane, a control (or management) plane and a network plane. The business plane enables access to the high-level administration databases and thus implements typical administration tasks. The entities in the control plane use these data when making decisions. The decisions are then enforced in the network plane using an AAA protocol and a policy-based management framework (PBM). The network plane itself is a pure IP environment providing users with network core services and service access. This is shown in Figure VI-2.

The basic problem with the SPN architecture is comparable to the one discussed before (see Section VI.1.5) but concerns higher layers. The existence of very different user groups with different expectations implies the provision of different services and service properties within the network plane. Since the network plane provides logical services (IP, i.e. L3 and higher), multiple network planes are not necessary. However, a multiplication of necessary instances for the diversified service provision should be minimized to achieve more efficiency. Instead, the number of instances should only depend on the maximum supported number of users. Therefore, mechanisms allowing a dynamic adaptation of the same instances to user demands and corresponding control entities have to be defined. Obviously, such control entities are involved in both user and SPN-internal inter-entity communications (e.g. controller to the controlled device).

The user-SPN relationship fulfilling the contractual promises over the chosen AN has to exist for any given user at the time of service access. Since networks are supposed to be different and the visited network is not necessarily the user's home network, this requires a dynamic adaptation of at least one of the two (user or network plane) and plays an important role in the logical access phase (see Section V.3.1). Therefore, the user-network interface has to be standardized to a certain degree. We have already studied these issues in Chapter V trying to minimize the necessary standardization.

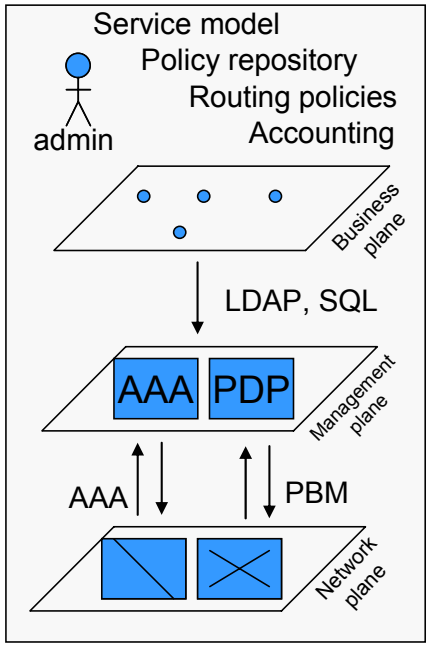


Figure VI-2 Service provider network with three major planes

In contrary, we believe that the internal SPN architecture should not be standardized. Indeed, not the implementation of services but their availability, quality and price are main drivers for user satisfaction. Providers should be interested in this additional degree of freedom: it gives them more flexibility in almost any aspect of their daily business. A tight standardization would homogenize the market offers in terms of provider equipment and competitive provider offers thus aligning the needs and the possibilities. It therefore dictates several aspects of the business model. Providers should keep their freedom in choosing what and how they want to organize in their network. Obviously, the complexity of this organization is related to the possibilities to provide more reliable and more complete offers to users. This can be easily expressed by different prices. In any case, the internal implementation should be hidden from users. This virtualization is necessary in this scope to guarantee the flexibility of every provider. Besides, this also corresponds to the user-oriented design which we pursue in this work since users' participation in technical issues should be avoided to achieve a better system ergonomics.

In the following sections, we present two different approaches to this problem. The first approach introduces a system featuring a user-parameterized adaptation of the common network plane of the visited SPN to the user profile. It is capable of providing different, personalized service environments for different users without implying too many constraints on the internal organization. The second approach shows how the terminal can be extended by a provider-issued module that can then adapt dynamically to a given visited SPN hiding the complexity of the internal organization. Finally, we give a short comparison of both approaches trying to identify their respective capabilities and limits.

VI.3. RESACO: User-Parameterized Adaptive Network

RESACO which stands for "RESeaux Adaptatifs et COoperatifs"[1] is a research project that studies dynamic network plane adaptation to the user preferences [128]. The objective of this project is to propose viable architectures and tools that enable and enhance QoS, mobility, service discovery and security aspects over multi-domain and heterogeneous networks. The proposed extensions are based on the use of programmable software-based edge routers and access points. We also experimented with other features that enable adaptability, auto-configuration and cooperation.

In particularly, we address service discovery, adaptive auto-configurable and programmable access and the inter domain security.

VI.3.1. "Virtual Network" Architecture

The basic idea behind RESACO is the provision of a programmable network plane. Such a dynamically programmable network plane is capable of adapting to multiple user requests "on-the-fly". Using profile data from the logical access phase and local policies, the network transforms to a service environment expected by the accessing user, as defined in the corresponding profile. This is applied to the whole network plane, including routers and services. It can involve AN adjustments. The goal is to provide users with the expected quality of the available services independently of the chosen SPN; "expected" means here that it corresponds to the user's contract. The network thus has to be capable of adapting to the user's profile providing a habitual environment. That is what we call a *virtual network* concept.

VI.3.2. Network Plane Adaptation

The SPN can thus be seen as a common base structure that can take different concrete shapes depending solely on the user profile and the current policy (with criteria like current load, SLA, etc.). However, users do not perceive the whole complexity of the policy-based decision mechanisms. From the user's point of view, the whole complexity comes down to a simple accept or reject of the access request. Hence, the active policy mainly remains an internal SPN consideration. In our view, the RESACO concept defines the user-experienced serving network (SN) as a result of the dynamic adaptation of the underlying SPN to the user's profile:

```
SN(SPN, U) = SPN.adapt(U)
```

where U – the user's profile. We study possible implementations of the `adapt()` function for RESACO in different contexts. The `adapt()` management function is part of the provider's internal implementation and considers internal policies and the SLA with the home provider of the user in question. This function can be proprietary. From the system point of view, only the user access thus has to be specified (interface specification as with every virtual concept). The user access part mainly concerns the access equipment (access admission, QoS), but can also involve higher level access to the control plane. The `adapt()` function is triggered in different SPN components (AN, routers, services, etc.) by SPN's access controllers as soon as the user profile is available. In this view, the

[1] in English: "Adaptive and Cooperative Networks"

`adapt()` function is a private function of the SPN class. RESACO does however propose and study the usage of MIT's CLICK language [129] as a possible implementation of `adapt()` in the core network routers.

VI.3.3. Profile-based Service Discovery

Main Idea

As we have already pointed out, a chosen AN typically changes the availability of services, since some services simply do not make much sense over certain technologies (e.g. data transfer over GSM). In such changing service environment, a service discovery procedure (see Section IV.4.4) and a corresponding SPN infrastructure seem indispensable in spite of SPN's adaptation capabilities. It also provides more flexibility: the `SPN.adapt()` function does not always result in the expected service environment. Discrepancies to a given profile could occur due to policy restrictions, failures, etc. A service discovery mechanism permits to detect and to react to such discrepancies. In this scope, we have studied and worked out practicable service discovery mechanisms.

To be able to adapt to different user profiles, the underlying SPN is likely to implement a superset of services expected by every individual user. However, logically the service discovery mechanism is situated at the user interface. It has to act in the resulting virtual `SN(SPN, U)`. In that user-profile parameterized network, a service discovery mechanism can only further restrict the set of the available services (as consequence/reaction to failures or policy decisions). Thus, the service discovery mechanism itself has to be parameterized by the user profile. Logically, it must not ever discover services not available in the profile.

Possible Implementation

In our model the SPN is based on the IP technology (see Section V.1.3). The service discovery infrastructure is logically situated in the control plane of the SPN. To enable exchanges of user equipment with the control plane, we use the standard approach that is based on the IP technology.

For reasons of availability [130], simplicity, media independence, operating system compatibility and no imposed device-API constraints we use UPnP (Universal Plug and Play) as service discovery framework. These properties make it especially interesting for various user devices. Besides, UPnP is primarily based on well-known Web technologies like TCP/IP, HTTP/SOAP and XML. It thus naturally fits the SPN IP environment. UPnP is a hybrid reactive/proactive protocol (see IV.4.4) and supports both service advertisements and active service searches. The UPnP framework defines two main entities:

- Control Points (CP), usually installed on the user equipment
- Devices which offer services

Control points search for devices (reactive) while Devices advertise (proactive) their services. The defined advertisement messages use multicast addresses. The advertisements are treated like events, these are thus not sent periodically but only when service state changes occur (internal state changes, expiration, etc.). User CPs can subscribe for certain events. Subscriptions and event advertisements permit to minimize the number of multicast messages and also to control to whom which messages are sent. In the UPnP framework, Devices can be further decomposed in one Root Device containing one or more Embedded Devices that in turn contain one or more Services.

According to the current specification [88], three advertisement messages are used for the root device itself, two per each installed embedded device and one per contained service. For k root devices with m embedded devices each and s services per embedded device it thus results in:

$$N_{mcast}(k,m,s) = 3k + 2km + kms$$

multicast messages per service advertisement. Obviously, the minimization of the number of devices minimizes the number of multicast messages for a given number of services. The best case is $N(1,1,s) = 6s$ messages per advertisement. The conclusion is thus that within UPnP it would be better to install as many services on the same device as possible. Unfortunately, in practice it is impossible to install all services on one device.

In RESACO, to enable a big number of possible $SN_{SPN}(U)$, s can be potentially high. Since control points are installed on the user equipment connected over a wireless link, this leads to an unacceptably high number of multicast messages on wireless links. To minimize the number of multicast messages on the user link, we introduce a new logical entity in the control plane of our SPN. This Network UPnP Interface (NUI) is a gateway between the SPN network plane services and the user equipment. NUI's primary function is to act as UPnP Device advertising services to the CPs installed on the user equipment and answering their search requests. To do that, NUI possesses the service status information on all available SPN services. This information can be obtained by proprietary means. For instance, NUI can additionally act as UPnP CP gathering information on all available SPN services, as shown in Figure VI-3. This however primarily depends on the expected dynamicity of the service environment and can be freely chosen by the provider. In some cases, NUI could simply extract such information from the internal databases. From the system point of view, the only mandatory part is the usage of UPnP on the user interface.

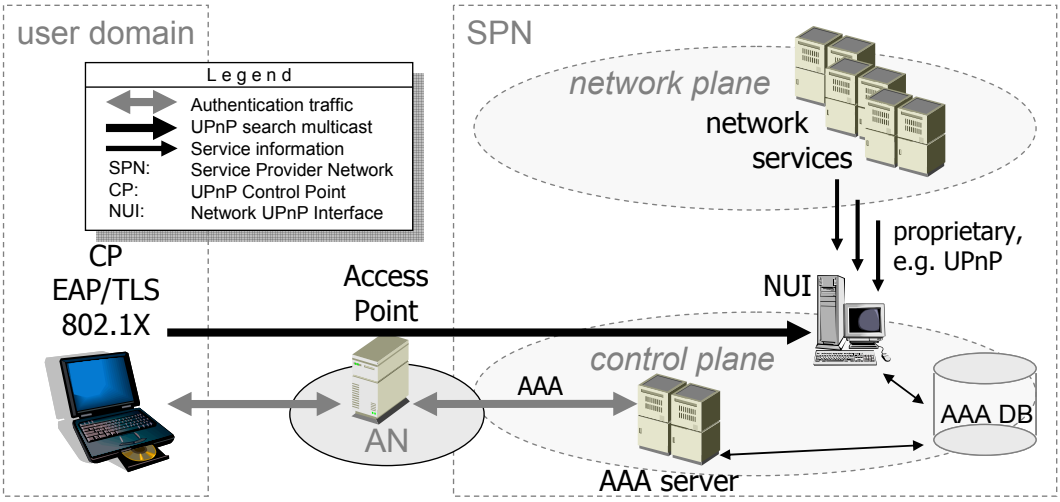


Figure VI-3 Access control and service discovery in RESACO

The NUI entity contains one root device logically representing the visited SPN and one embedded device logically representing the used AN (e.g. AP). NUI contains information on s services. The introduction of NUI permits to achieve higher efficiency:

- NUI permits to minimize the number of multicast messages exchanged over the wireless user link without enforcing particular service installation conditions.

- NUI minimizes the service discovery latency on user access since it already possesses all necessary service data (description, configuration, information, etc.) when the user accesses the network.

However, NUI also opens new opportunities:

- By decoupling the internal network service discovery from the user link service discovery, it combines the compatibility (indispensable for users) with the flexibility of the internal implementation (essential for the providers). Indeed, the constraints and the requirements on the service discovery mechanisms can be very different for the internal use and the user link.
- As a control plane entity, NUI can filter its discovery replies depending on the user identity effectively implementing profile-based service discovery which we have motivated above. That is also true for NUI initiated advertisements sent to the multicast address: first, we can separate users in the IP address space; second, UPnP uses a subscription-based advertisement mechanism and NUI can reject subscriptions if the service is not available or not allowed according to the profile.

NUI is a logical entity providing an efficient user interface for service discovery. It uses the well-known and wide spread UPnP for this purpose. Note that we make no assumptions on the actual implementation of NUI. In particular, NUI is not necessarily implemented as a central entity. For instance, in our test environment, NUI is completely distributed, installed directly in the access points.

VI.3.4. Profile-Based Service Access Authorization

The profile-based service discovery avoids detection and configuration of services that are not available in the user profile. Obviously, this has to be further enforced by service access control measures. We distinguish two different service access control measures.

The security measures are capable of qualitative (is it allowed?) admissions and block any access to an unavailable service (in the $SN_{SPN}(U)$). Quantitative admissions are necessary to enforce/guarantee a service quality conforming to the service contract. The latter should be bilateral: user traffic has to be observed, controlled and made to conform the contract while the service performance must be continuously measured and adapted to fulfill the service contract.

In both cases, the user profile contains the information necessary for the parameterization of the access mechanisms.

Since services in the SPN are IP-based, service access is also based on IP. We thus use profile-parameterized dynamic packet filters to qualitatively block out any unwanted access. On every permitted connection we apply IP-based counters and QoS filters that then control and enforce the necessary QoS parameters. In our test platform, we used a restrictive filtering policy (what is not explicitly allowed, is prohibited). We applied the IP filters directly in the first network equipment: the access point.

VI.3.5. Control and Enforcement in the Edge Equipment

RESACO SPN architecture follows the paradigm of early policy enforcement. The main idea behind this architecture is to achieve the maximum possible simplification of the network plane, transforming it into a mere QoS-aware IP switching fabric. The network plane thus approaches the common IP core in the 4G architecture.

To achieve this, we push the control elements away from the network plane core closest to the user, i.e. in the edge equipment. This follows the best practices developed in the ATM and MPLS research. In our case, the equipment closest to the user is the access point of the access network establishing the first link. The integrated control points comprise L2 and L3 security controllers, service discovery interface, packet filters, traffic shapers, etc. This early control (from the first link on) assures that only conforming user traffic can enter the network plane from the user domain, effectively achieving the simplification of the latter. It achieves higher security preventing non-conforming user traffic from entering deep into the visited SPN. Besides, this approach scales better since we presume that the number of the deployed access points automatically follows the number of potential users.

The control plane includes all decision points and the associated databases. In our architecture, control entities can be installed in a small, clearly separated network part. This design increases the transparency of the architecture. It also allows precise control of the incoming and outgoing control traffic. Indeed, in our architecture the only entities communicating with the control plane are trusted SPN equipment entities like access points from the edge, routers from the network plane and services. Thus, strong reliable security measures are possible, based on the existing security associations. These measures protect the edge equipment and the programmable network plane core from internal attacks. Moreover, the configuration of the separating firewalls can be static and thus particularly easy.

The business plane is implemented as a set of views on the control plane databases. The administrators can change stored policies and user and equipment profiles according to the administration goals.

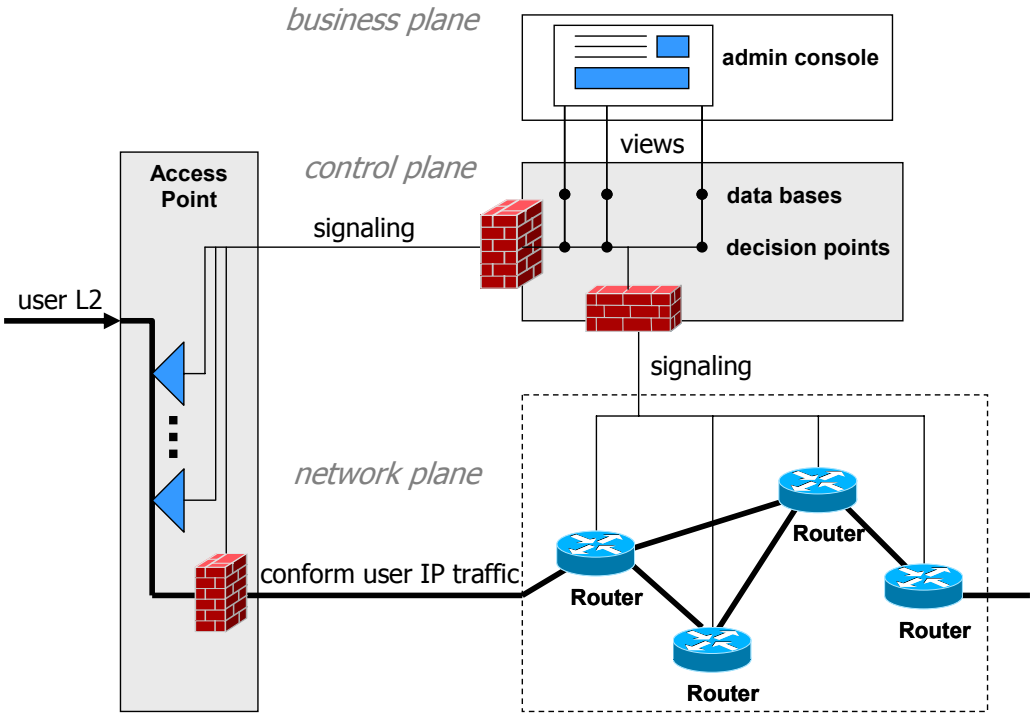


Figure VI-4 Main principle of the RESACO network architecture

VI.3.6. Test Environment Implementation

Our prototype implementation consists of an extended access point implementing the necessary control plane functions and a minimal architecture reflecting a real SPN environment. This is shown in Figure VI-5.

Control plane elements in our test bed include an integrated DIAMETER/PDP server for inter-domain QoS enforcement. This AAA server also builds the local controller part used for the 802.1X access control (over RADIUS). Both use an associated database for logging, current state repository and long time settings (like e.g. standard policies and user profiles). A web server permits access to these databases from the business plane.

The business plane views are implemented by using modern Web technology. We use dynamic Web sites on a Web server to create dynamic views of the databases. Administrators can access this interface from virtually everywhere with an SSL capable modern web browser. See Appendix E for the details of the implementation and the used software.

Our test SPN has four different services and the test environment uses four different user groups. The group authorizations to services are shown in Table VI-3. We use Squid, an open source web proxy. It is known for its reliability and is often used in commercial systems. The multimedia services use software from the VideoLAN project as both sinks and sources. VideoLAN can dynamically adapt its quality to the properties of the underlying network. In the test environment, both sources and the web proxy are installed on the same host. This is not necessary.

Table VI-3 User groups, services and authorizations in the test environment

	802.11b access	Web proxy	Audio service	Video stream
Unauthorized	no access	no access	no access	no access
Visitors	Restricted 1Mbps	Yes	No	No
User group 1	Full	No	CBR, 128kbps	CBR, 300kbps
User group 2	Full	No	No	CBR, 2Mbps

The proposed solutions for open and adaptive WLAN access points are based on extensions of the HostAP project. We use Linux as the host OS for the HostAP driver. In order to enable end-to-end harmonized QoS over the wireless interface and the wired transport network, the first extensions are related to QoS handling. A QoS differentiation feature is added on the radio interface part of HostAP. DiffServ routing capabilities are added on the transport interface. A dynamic Policy Based Management framework (PBM) [131], based on the Common Open Policy Service (COPS) protocol [85], has been integrated into this open and programmable access point. This PBM allows the dynamic management and configuration of both wireless and fixed interfaces for both uplink and downlink flows. All key parameters related to resource management for both radio and fixed networks interfaces are stored into dedicated Management Information Bases (MIBs). HostAP also integrates a RADIUS client to demand remote authentication at an authentication server.

The two main components of the PBM framework are the Policy Decision Point (PDP) responsible for central decision-making and the Policy Enforcement Point (PEP) responsible for applying these decisions on the managed network entities. A Policy Repository (Policy Rep), containing policies and rules to be applied, is attached to the PDP. The COPS protocol is used to exchange management information and decisions.

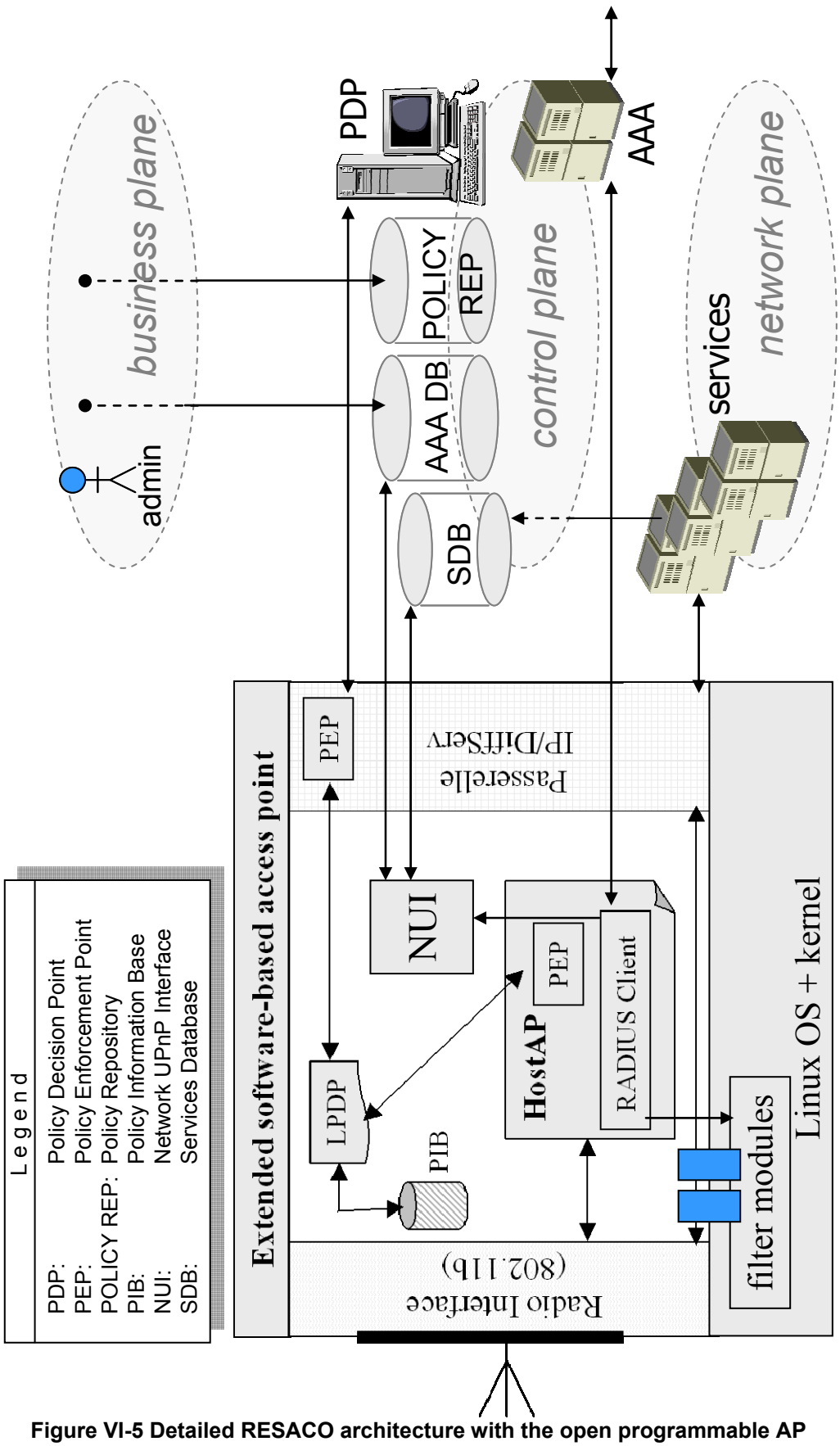


Figure VI-5 Detailed RESACO architecture with the open programmable AP

In the proposed architecture, a local PDP (LPDP) and PEP entities are added to the AP:

- The PEP entity integrated in HostAP is responsible for enforcing the decisions concerning the management of the radio part and its dynamic bandwidth allocation in the `iptables` filtering module (netfilter).
- A local PDP (LPDP) and its corresponding PIB is included in the architecture to make local decisions and minimize data exchange with the remote PDP.

PEP instances are also integrated in the DiffServ routing module of all routers in the network plane to enforce the decisions concerning DiffServ parameters and policies on the fixed network segment.

The inter-domain roaming access follows the architecture proposed and studied in the AUTHENTIS project (see Section V.4). The RESACO demonstration environment uses an IEEE 802.11b access network. We use profile-based access control by the means of IEEE 802.1X mechanism. The AUTHENTIS mechanisms are principally suitable for user profile delivery to the visiting network in the 4G scope. The remote authorization concept has been extended to carry inter-domain QoS negotiation (e.g. for home service access) Based on the DiffServ approach, it uses DIAMETER. More details on this are in [132].

Service discovery is implemented by integrating the introduced NUI entity with the RADIUS-based access control. In case of successful authentication, the incoming `Access-Accept` message is used in the HostAP's RADIUS client to trigger an instance of the previously introduced NUI entity. The latter uses the AAA DB and the available service information (stored in the Service Database, SDB) to determine which services need to be advertised to the user in question. The service set to be advertised to a user can be obtained by disjunction of the service set available in the local SDB with the allowed service set in the user profile.

Service access control and parameter enforcement are implemented by using the `iptables` interface to the filtering capabilities of the Linux 2.4 kernel. `iptables` system calls are effectuated per user on user access by the modified RADIUS client.

In RESACO we do not study the syntax and semantics issues related to policies and service definitions. This problem is generally unresolved (meaning that there is no common and consistent language for service naming or for policy definitions). Technically, AAA attribute-value pairs (AVP) or descriptions in suitable XML variants can be used for these purposes. We think however that the best solution would avoid monstrous generalized definition languages since these tend to exhibit a very complicated grammar. In the best case, the naming schemes and policy languages should remain provider-intern. The mutual mappings can be agreed upon in the SLA, thus defining pairwise mapping functions.

VI.3.7. Results

In our prototype implementation we demonstrate [133] that two users from different user groups experience a different network within the same deployed SPN/AN infrastructure. More precisely, we verify the following points:

- Unauthorized users are denied access to all services
- Visitors can get a restricted Internet access over the available Web proxy
- The discovered service sets are different qualitatively (service lists per se) and quantitatively (properties of the included services). Among the discovered data there is a full, medium-independent SPN name and the name of the visited subnet (see Figure VI-6).

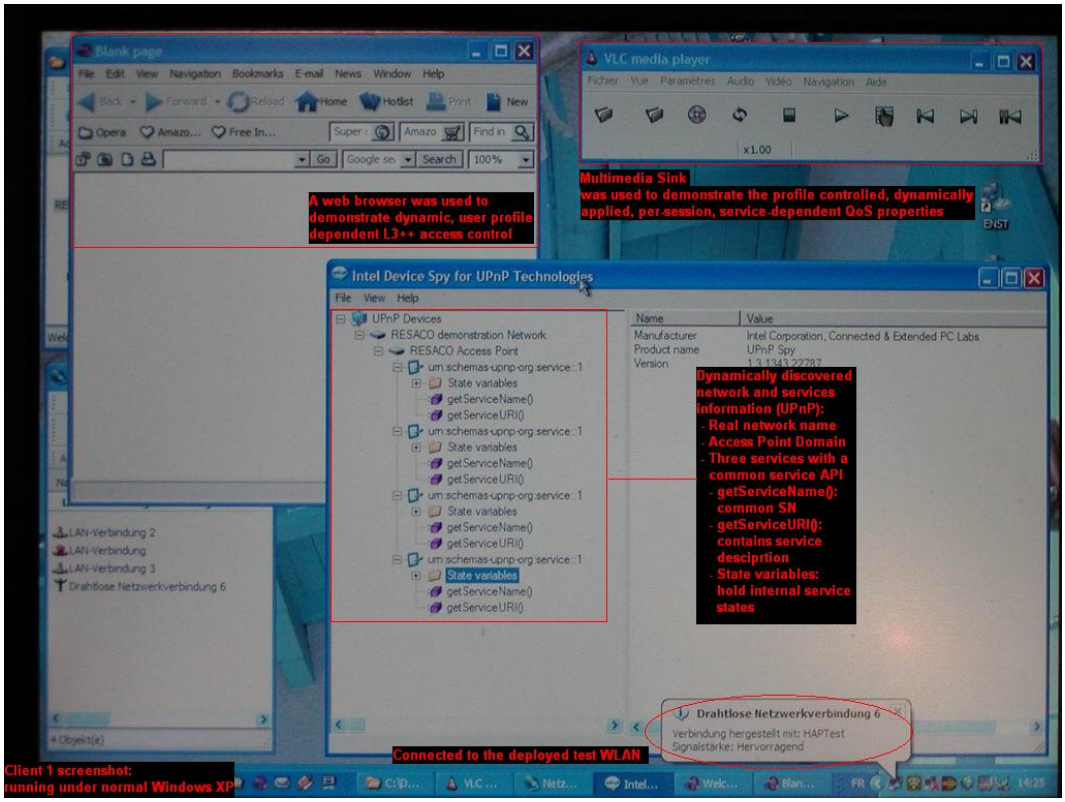


Figure VI-6 User from group 1 connected to the demonstration network

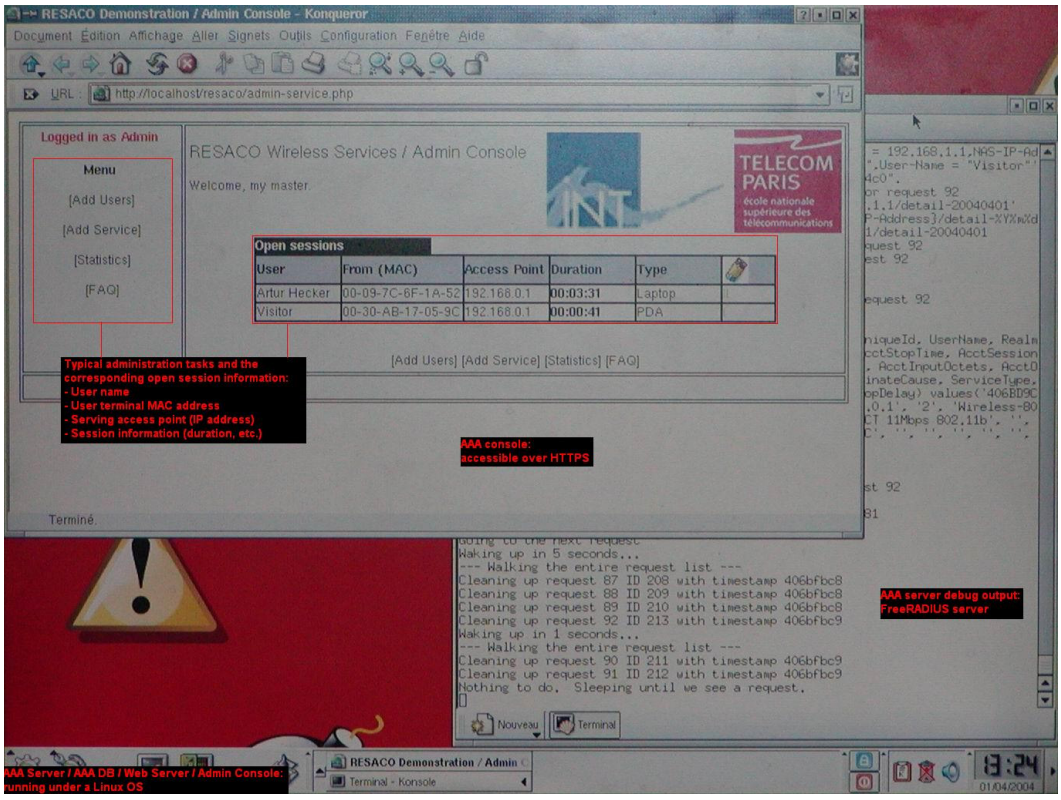


Figure VI-7 Administration console in our test environment (two connected users)

- The discovered service data are organized in form of a common API that can be used to fetch further data on service configuration and location, internal service variables and humanly readable service description URIs (Figure VI-6).
- We have shown that for both user groups the access to services is qualitatively and quantitatively different.
- Network administrators are capable of following and of influencing the situation in the deployed network by using our management interface (see Figure VI-7).

These and other images from the demonstration in Brest can be downloaded under [128].

VI.3.8. Conclusion

Within RESACO, we study provider networks architectures capable of adaptation to the profile of the accessing user.

We have identified several requirements to provide users with a homogeneous experience over heterogeneous networks. We proposed and developed viable mechanisms for profile-based service discovery and QoS-aware service access control in the future IP-based service provider networks. We have pursued the approach of pushing control elements away from the purely IP-based core, as close as possible to the user equipment. We propose access equipment fulfilling the necessary tasks like per-user L2 and L3 access control, inter-domain user profile fetch, profile-based service discovery and roaming service access with QoS enforcements.

We integrate the most control elements in the first encountered network component, the access point. We think that with the ongoing development of the integrated circuits and the associated steady price decrease such integrations are already possible for the industry to competitive prices. Our working prototype shows that such integration is also feasible for “cheap” networks like 802.11. To reduce the development costs, in our platform we have integrated different software bricks available as general public license (GPL) or open source software. We then used an old notebook with a Linux OS as host for the prototype. Our tests have not revealed any performance bottlenecks. In fact, the system has never approached its performance limits in terms of CPU or memory requirements. Indeed, our design does not require complex per-packet treatment, since most tasks (except packet filtering) are carried out during the network access phase.

The recent release of low cost 802.11 access points using embedded Linux [151] and integrating different software bricks including an embedded Linux kernel and the kernel-based packet filtering prove that our approach is viable for the embedded OS on the current AP platforms.

VI.4. MMQOS: Virtualization of Services by Using Terminal Smartcards

We pursue our basic principle of providing users with a homogeneous service environment over heterogeneous networks, preserving the independence of the provider-own infrastructures and mechanisms. In this section, we present an alternative approach to the adaptation of the user-network interface.

This work has been carried out within the French national research project in telecommunications (RNRT) named MMQOS [134]. Our participation in this project

included a system security analysis, the proposal of various alternative access models and the definition of the corresponding system architecture.

VI.4.1. Virtual Network Interface

In the last section we have defined an adaptation function in the visited SPN. In this part, we define an adaptation on the other side of the link. We show how the terminal can be adapted to a visited network without coupling the internal network implementation to the user link. This apparent contradiction can be resolved by using an active home provider module within the user terminal. This module implements the mechanisms corresponding to the provider SPN organization and hides these from the user. Therefore, users can obtain a homogenized access over a virtual network interface provided by this module.

The active module has to be responsible for security, QoS, service access and other issues which we have already discussed above. Since it is involved in the security-relevant provider-internal measures, a protected execution environment is considered in this work. We propose to use smartcards for this purpose as shown in Figure VI-8.

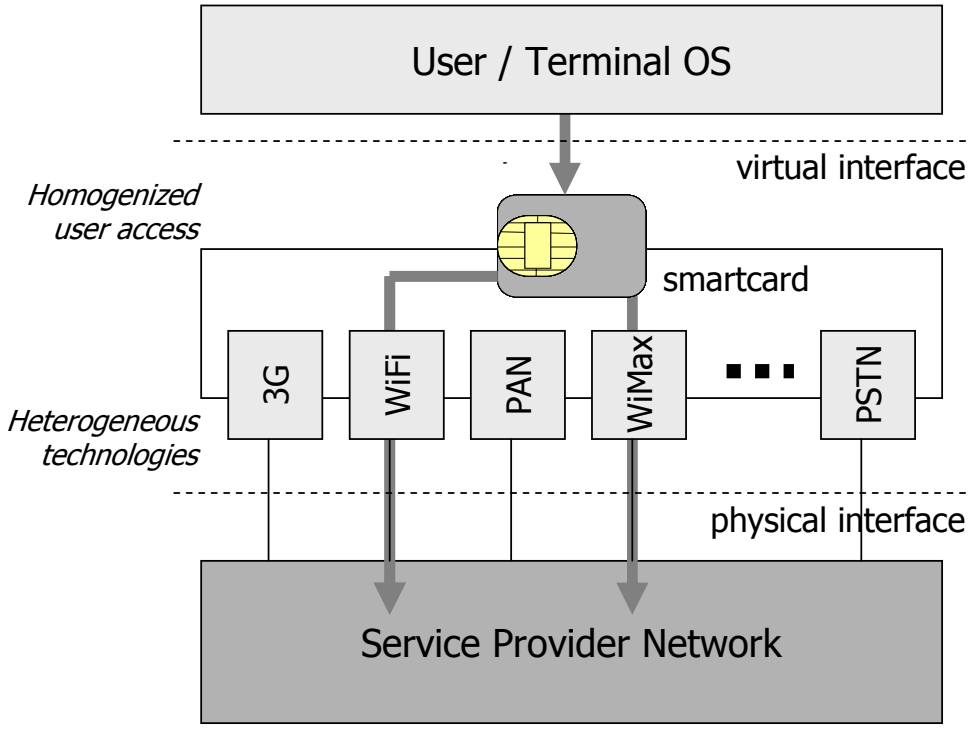


Figure VI-8 Virtual interface provided by the smartcard in the terminal

VI.4.2. IP Smartcard

Description

In this proposal, we pursue the strategy of pushing the control elements further towards the network edge. In fact, the smartcard installed in the user terminal acts as a network edge device. This smartcard is responsible for different control and network layer tasks. However, our network plane is completely IP-oriented and different control layer tasks can be IP-based. The smartcard thus has to be accessible over IP since otherwise the used mechanisms and protocols have to be adapted.

TCP/IP-enabled smartcards like the SIM-IP card [135] have recently gained popularity in the industry. Different manufacturers are planning to propose smartcards integrating a TCP/IP stack [136]. We think that this trend will gain momentum with the proliferation of IP technology. Furthermore, the smartcard development process produces faster and more powerful devices every year. In this work, we use the SIM-IP card as an example because it is the first card to propose an integrated TCP/IP stack.

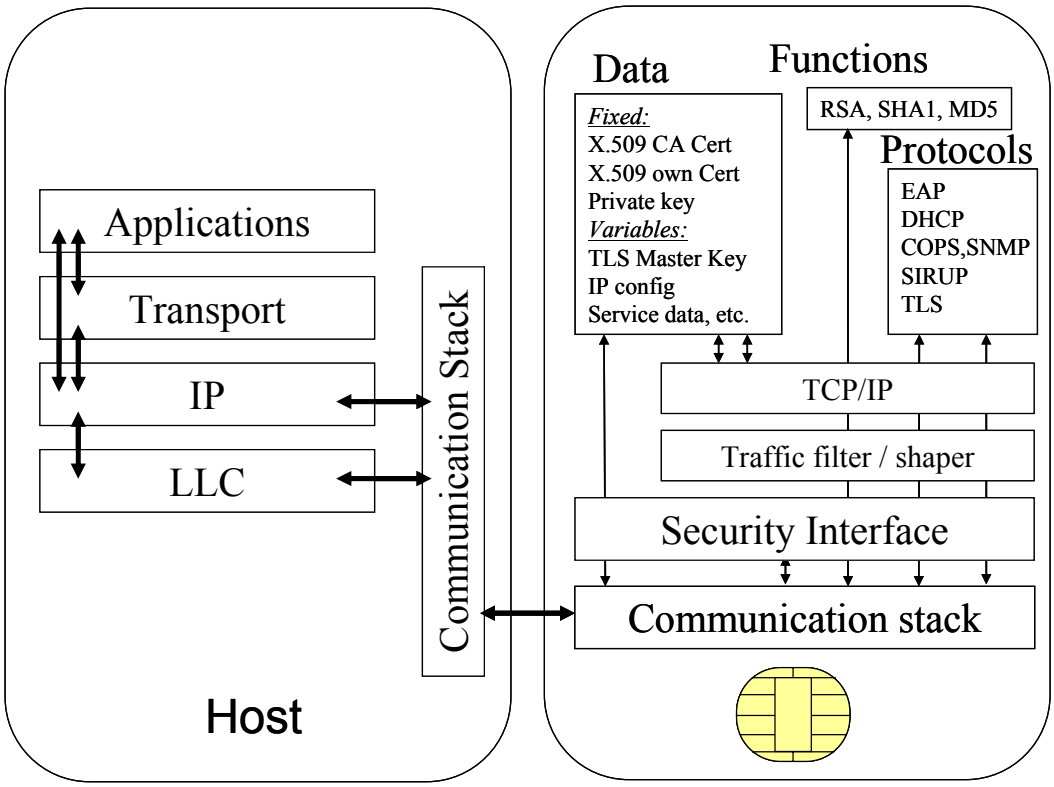


Figure VI-9 SIM-IP card

The SIM-IP card is an IP-capable Java-based intelligent subscriber's identity module (SIM) with the possibility to integrate services (see Figure VI-9). Similarly to the GSM/3G (U)SIM cards, it provides a mechanism for user authentication and accounting. Additionally, it includes TCP/IP functionality and can, on one hand, complete terminals that do not support IP natively and, on the other hand, provide IP-based services itself. The card carries a set of security associations (user credentials, authentication procedures, algorithms, etc). It stores data in protected XML files. It can execute Java applets in a protected environment. In particular, it integrates a highly trusted web server and supports various protocols like HTTP, LDAP, COPS/SNMP, EAP, etc.

Additionally to that trusted and tamper-resistant computing environment, the SIM-IP card offers three main advantages:

- Abstraction from the technology-bound secure access mechanisms of the used access network technology
- TCP/IP stack independent from the associated terminal
- An opportunity to include service end points on the card

We use these SIM-IP features to provide users with a computing environment independent of the serving network. For that purpose, we introduce a novel Services-on-Card (SoC) concept. Since we install classically network-internal components on the

card, each SIM-IP remains the property of the issuing provider. It is pre-configured by the latter and seen as a trusted network node after it has successfully established the link. This is illustrated in Figure VI-10. Using trust transitivity (see Section V.2.3), this can be easily extended by an additional visited network. Note that the SIM-IP card does not implement any radio access specific functions.

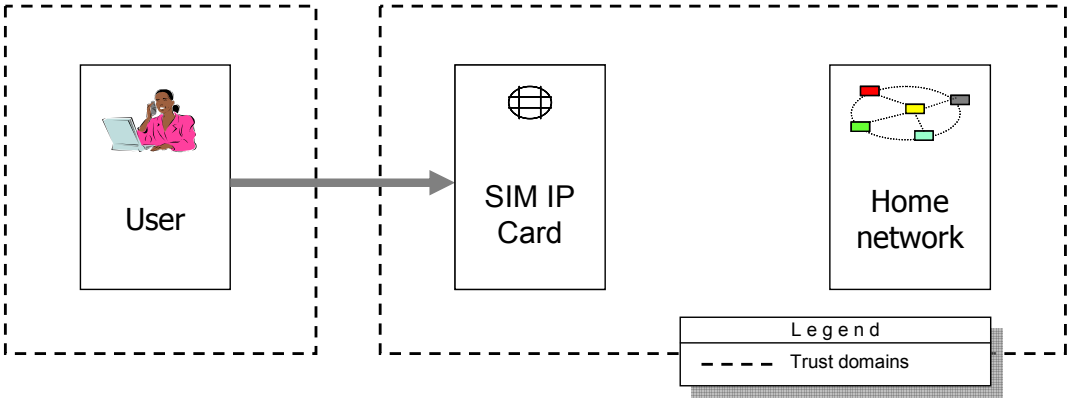


Figure VI-10 Main actors and trust with the SIM-IP card

We distinguish four *network access phases*. In the first phase, the card connects to all available SPNs using contained credentials and algorithms over terminal's network interfaces. In the second phase, the card collects the data necessary for the network selection decision. In the third phase, the card verifies the user credentials, presents to the user the available services and their properties and grants service access to users. Herein, the user verification is very simple since it can be processed internally by some proprietary algorithm (typically, smartcards use PINs or passwords). Even the mere possession of the card might suffice in some cases.

Finally in the last phase, after a successful network access, the card classifies and manages user traffic and makes necessary reservations. A QoS-aware traffic filter can be installed on the SIM-IP card. It classifies the passing IP packets according to the used protocol/application. Different traffic classes can be defined describing diverse criteria (throughput, latency, jitter, …). Each class is given a priority level.

The SoC concept and the phases are described in the following sections in more details.

SoC concept

The integrated logic of the card and its mentioned IP capabilities enable service prolongation from the network to the card. This presents new opportunities. An issuing provider is able to install some traffic control components such as classifiers, filters, packet counters, etc. on the card, thus allowing control or even enforcement of contract-consistent user behavior. Furthermore, a provider could integrate service access points in the SIM-IP card, thus offering the user what we call Services-on-Card (SoC), i.e. services available directly from the card, independently of the actual location of the related service end points. After the connection phase, the service access points located on the card dynamically choose the service provider (e.g. home vs. foreign network) depending on the availability of the service, service properties, etc. or interactively.

The network access service already represents a network-internal service prolonged to the SIM-IP card. From the user's point of view, the card acts as a network edge device granting or rejecting access. The control and enforcement points for QoS, mobility management and security mechanisms belong in the same category of core services. The

other category contains IP services like HTTP, SMTP or SIP. Such services can be implemented as proxies configured by default to contact the home network. The card issuer could assure the service availability in the own network and design the card according to the available offer. In the case of service presence in the visited network, the SoC-proxy can be dynamically reconfigured by the card logic and use the local service provider. This is controlled by policies, service level agreements and user choice.

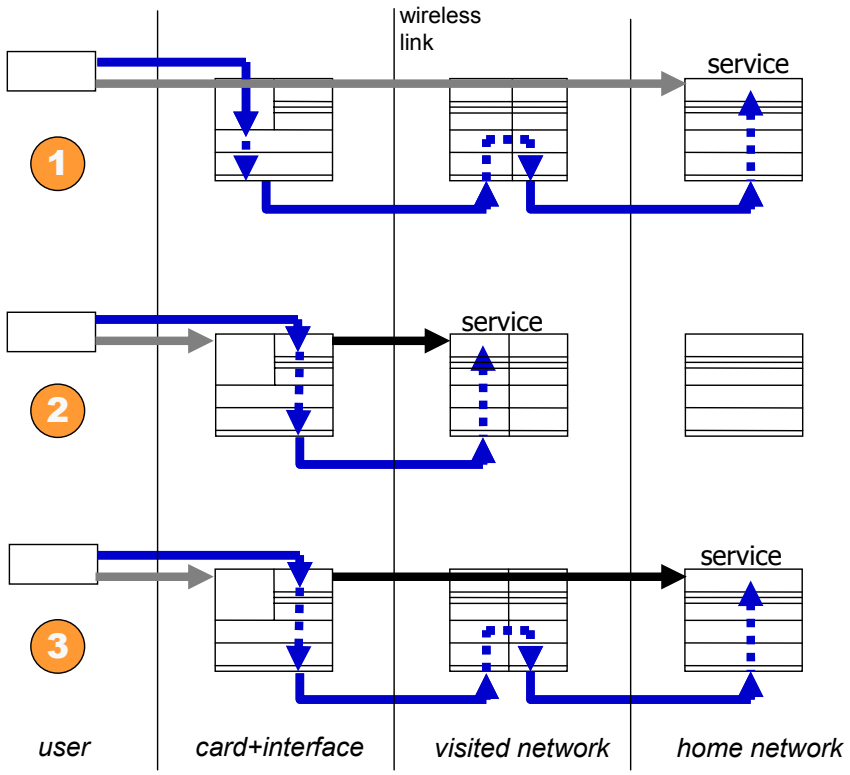


Figure VI-11 The SoC concept with the concerned OSI layers in three scenarios

The usage of the SoC (in terms of implied OSI layers, etc.) compared to the direct service usage is illustrated in Figure VI-11. It shows three different scenarios:

1. direct access to a service in the home SPN
2. access to a service in the visited SPN over a SoC proxy
3. access to a service in the home SPN over a SoC proxy

It can be seen that in cases with SoC usage (2 and 3), the user always accesses the card (grey arrow).

SoC & SIM-IP: Parallels to GSM SIM and Virtual Private Networks

Hence, the SIM-IP usage in that scope can be seen as a generalization of the GSM SIM concept. Indeed, GSM SIM hides the network access algorithms from both users and partner networks and thus transparently provides a security level required by the issuing provider for roaming users. SIM-IP principally does the same in the described heterogeneous technology context (very different security requirements, functions and mechanisms).

However, with SoC support SIM-IP goes beyond that security abstraction. The added value of this approach is the stability of the computing environment from the user's point

of view (no need to reconfigure the applications, no need for user actions, etc). In that sense, SIM-IP card resembles a virtual private network (VPN) module, widely used today by mobile users to securely access internal enterprise networks, Intranets, etc. However, as opposed to the VPN approach, SIM-IP permits to achieve this stability in the most efficient way, i.e. using the next possible service end point. Moreover, since it is done dynamically, it achieves a higher service availability. As with VPNs, the access to home services can also be provided in a secure way, completely transparently for the user and the visited network. This discussion is summarized in Table VI-4.

Table VI-4 Qualitative comparison between VPN, GSM SIM and SIM-IP approaches

	VPN	GSM SIM	SoC / SIM-IP
Security abstraction	Yes	Yes (for GSM)	Yes, also over heterogeneous links
Service availability	Only home services	Only local service	Both, dynamically chosen
Optimal service access	No (always remote)	Yes (always local)	Yes
Secure home network access	Yes	No	Possible

But is this approach always reasonable? Indeed, in the case where the service sets in all SPNs are equal, users can principally always access services locally, in the visited SPN (e.g. with GSM roaming). In this case, the effort is almost useless (the SoC concept could still provide more robustness). In the case where the service sets are strictly disjoint, visited SPN services are unknown to the card and the home service counterparts are not available in the visited network. It results in the constant connection to the home network (like with today's VPN modules). Hence, to provide the advantage of the SoC-idea, the common service set should be a real superset of the core services but a real subset of the result of the disjunction of all service sets. In the open environment with competing providers of different sizes and commercial orientations, this seems quite likely.

Implications

The installation of the SIM-IP card on a terminal requires initial non-recurring changes in the system and user application configurations. This is similar to the VPN approach.

Obviously, for the card itself we have to provide a discovery and dynamic configuration mechanism. During or after the card authentication phase, the card has to find out which services are available in the visited SPN. Then, for every potential SoC it is reasonable to reconfigure the on-card proxy respectively (e.g. configure the on-card SMTP-smarthost to point to the SMTP-relay available in the currently visited network).

Most importantly, we must not forget the implications on the security when placing potentially network-internal elements on the card. Usually, such "network-internal" protocols are not sufficiently secured by the protocol design itself. To prevent attacks, such protocols rely on external protection measures such as underlying encryption (e.g. by IPSec, secure tunneling, etc.) or an appropriate network design (e.g. architectural separation of user and management traffic). In case of installation of network-internal components on the SIM-IP card, such protocols are used on the wireless user link, in parallel to user traffic. Obviously, we have to secure such management traffic against possible fraud. Since these protocols are used after a successful SIM-IP card connection to the visited network, the related packets traverse the distance from the card to the secure core network over a link secured by L2 security mechanisms. In this manner, assuming a reliable confidentiality function (e.g. strong encryption), no third party can read the data emitted by the AP or by the card. If L2 measures are unreliable or insufficient, other protection measures should be conceived (TLS, VPN technology, IPSec, etc).

Control Entities on Card

In our general architecture, the PDP is a central SPN entity, while PEPs are installed in various devices, e.g. at the network edge. However, in the context of wireless access links this has one main disadvantage: the users can still behave incorrectly on the link between the terminal equipment and the edge device, i.e. exactly on the link with limited shared resources.

We propose to install a PEP on the SIM-IP card (see Figure VI-9). This gives new opportunities like enforcing QoS policies depending on the link load and controlling user traffic at its source preventing the OS (user) from sending out-of-contract packets. This permits load optimization on wireless links.

Card's Key Management

In our proposal, the card is the primary key manager for all user access possibilities. This includes the network access procedure, the network core service access, general service access and, optionally, direct home network access that might be necessary for some services (e.g. for a home-VPN key derivation module, if available).

Card's internal key management has been studied in the MMQOS project and published e.g. in [137]. As already mentioned, we demand that any suitable L2 authentication procedure be able to establish independent session keys (in the best case, with PFS – see Section IV.4.1). The basic idea is to minimize the number of statically stored keys and to propose a derivation and delivery scheme taking into account the L2 authentication procedure and the possible access to the services in both the visited and the home SPNs.

VI.4.3. Service Provider Network Architecture

The proposed SPN architecture follows the three plane view defined above (see Figure VI-2). The administration data can be stored in an LDAP directory in form of suitable Authorization, Authentication and Accounting (AAA) attribute value pairs (AVP) or COPS/SNMP objects. As in RESACO, an AAA protocol and a management framework are used for access control and policy-based management. RADIUS is principally sufficient as an AAA framework (since we do not use server-to-client push with AAA), but obviously DIAMETER can be used as well. Although we use the COPS nomenclature distinguishing the Policy Decision and Policy Enforcement Points (PDP/PEP), we do not specify which protocol is used for the management tasks. COPS or SNMP are suitable candidates. Proprietary methods (e.g. an automated console- or web-based management) can also be used. That corresponds to our policy to give providers additional freedom. However, we show how specific management tasks can be implemented at examples.

Figure VI-12 illustrates the SPN architecture in more detail. The user is represented by the operating system (OS) that executes user commands. The OS uses the SIM-IP card to access the actual terminal equipment (TER), i.e. generally the network adapter. The network itself is characterized by an edge device i.e. an access point (AP), the entities responsible for the core services like access control, management, QoS support, etc. (AAA, PDP, SessionDB, policy repository, user DB) and some optional user servers that depend on the provider. The core services are described below as part of the actual architecture. On the user side, the SIM-IP card implements mechanisms necessary for the core services. Core services include:

- Network access control (based on L2 access control of the used technology)
- Dynamic QoS provision (based on a suitable management protocol, see above)
- Service access (based on the IP identity within sent packets)

Every provider claiming to support our solution has to integrate at least the necessary network access and QoS management services as presented here. After reconnection, SIM-IP card acts as a part of the visited network (both instances trust the home provider and thus can establish a local trust relationship).

Hence, first the SIM-IP card connects to the visited (home or foreign) network. Obviously, due to the diversity of the potential physical networks, the SIM-IP card has to support the logical network access methods specified within each used technology. The card has to be able to answer to the challenges of the edge devices of the visited networks, e.g. access points or base station transceivers, since it is the only equipment that carries the needed security associations (credentials and algorithms). Therefore, the first link, i.e. the wireless link between TER and AP is protected by OSI layer 2 (L2) security mechanisms based on card's security associations. In some later phase, the SIM-IP card verifies the user identity.

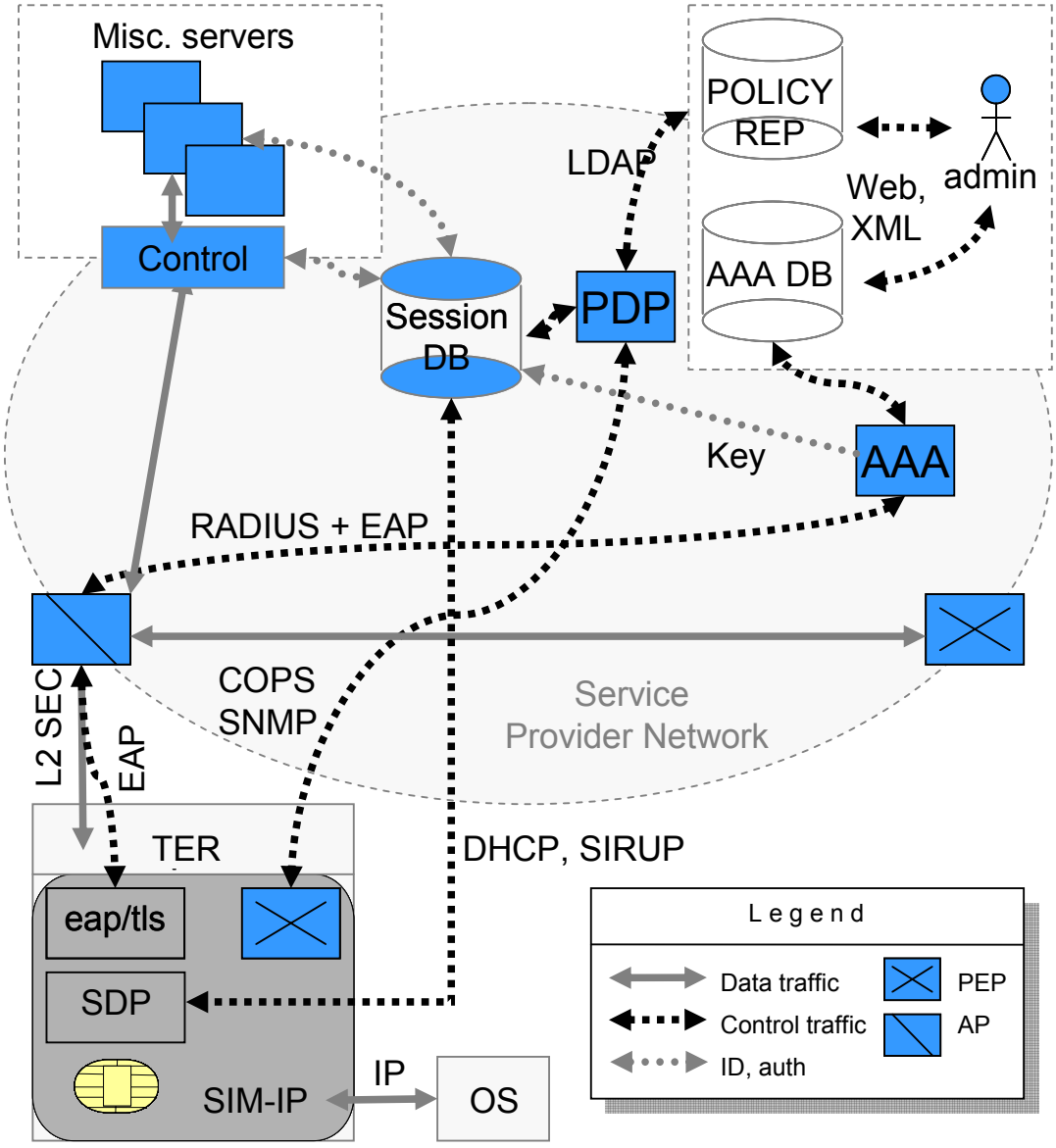


Figure VI-12 Service provider architecture

Unfortunately, the L2 security mechanisms do not solve the problem of later user identification for service access since IP-based services in the SPN do not naturally have the L2-information at their disposal to determine user's identity. We address this issue in the next section.

We assume that the cores of the visited networks are internally secure or can be secured by the responsible authority by means of physical subnet separation, encryption or other well-known, appropriate measures. The infrastructure security thus addresses the network access security and protection from an unauthorized resource usage.

User management data are stored in the databases. We mainly distinguish:

- A policy repository maintaining all needed policies (Policy Rep)
- An AAA database (AAA DB) as a part of authentication and authorization infrastructure
- A session database (SessionDB) to collect all data on users connected to the SPN

As illustrated in Figure VI-12, AAA, SIM-IP card, PBM entities and the SessionDB are the key elements of this architecture.

VI.4.4. Network Access Architecture

In this section, we describe how a user securely accesses a visited SPN. We discuss the associated architecture and explicitly cover roaming support.

Card to Network Connection

SIM-IP card stores the necessary security associations and implements the protocols and algorithms necessary for authentication, key management, etc. of every used technology The link layer security mechanisms like e.g. link encryption are carried out by the terminal interface itself. In the authentication phase, the SIM-IP card thus establishes the necessary key material. It then derives appropriate keys and delivers these to the selected terminal interface.

We discuss this at the example of 802.11 WLANs. We use the IEEE 802.1X architecture with EAP-TLS. The necessary security association is defined by TLS [77]. In our proposal, the card's private key, the certificate (signed public key) and the certification authority's (CA) certificate are stored in the SIM-IP card (see Figure VI-9). The SIM-IP and the network's RADIUS-server authenticate mutually over EAP-TLS and negotiate a TLS master secret [77]. Both sides then independently derive the communication keys following procedures defined e.g. in the IEEE 802.11i draft. The RADIUS server hands out the pairwise key to the solicited network AP together with the confirmation of the successful authentication. The AP opens the associated communication port and activates security mechanisms on the concerned port using a dynamic link layer key created from the received key. Similarly, on the card's side, the SIM-IP card creates and installs the link layer keys in the network adapter and the OS then activates link layer encryption on the link. In this manner, the data transmission over the wireless link is first possible after the card and the network have mutually verified their identities. Moreover, it is protected (encrypted, signed, integrity-protected, etc. – depending on the link layer capacities) using highly dynamic, per-session keys. The session-duration is expected to depend on the estimated strength of the L2 security. For current 802.11 WLANs we recommend a maximum allowed session of about 1-10 min. This should be configured in the corresponding card profile. The AP and the network adapter can and should change the used link keys frequently, i.e. also during the session. The mechanisms for this are part of the L2 definition. Using 802.1X, the card can always trigger a renegotiation.

Note that this phase describes the card's access to the network as opposed to user access to the card. The stored credentials are thus the card's credentials. These are logically independent of the user credentials. However, every card represents a service contract which, in our system model, is the base for the very user definition. Every card thus represents exactly one user. The home provider can correlate the latter with a user profile. Also, from the view of the visited SPN, the card is acting on the user's behalf and thus represents a contract-conform user terminal.

Roaming

To achieve the contract conformity, the SIM-IP card verifies user credentials, controls user traffic and grants users access to services. Since user always connects to and through the same SIM-IP card, no provisions for user roaming are necessary in our architecture. After the card's reconnection, the user can simply use the services in the same manner, independently of the visited network.

Conversely, we have to provide mechanisms for the SIM-IP card roaming. More precisely, there are three issues related to that problem:

- roaming network access,
- roaming SIM-IP card configuration,
- roaming service access.

These are addressed in the three following sections.

Roaming Network Access

The IEEE 802.1X standard proposed here for card network access recommends RADIUS usage as the backend authentication server. The roaming access with AAA/802.1X has been discussed in details in Section V.4.

The used 802.1X EAP-method (EAP-TLS) requires certification authorities and certificate deployment. However, a common Public Key Infrastructure (PKI) is not necessary i.e. every provider can simply install and maintain an own, independent certification authority (CA). Besides, our approach implicitly covers the private key and certificate deployment that is usually problematic when using EAP-TLS. In our approach, the private keys are installed in the card and never leave it. Certificates are stored in the XML files, publicly readable but not changeable. The provider installs these data on the card on card issuing.

We implement card roaming using a RADIUS feature called proxying [98]. The EAP-TLS conversation takes place between the SIM-IP card and the home network's RADIUS-server going over the AP and the visited network's RADIUS-server. The home RADIUS-server delivers the session keys to the foreign RADIUS server that hands it out to the concerned AP. The necessary provision for RADIUS proxying is the RADIUS-server interconnection of the concerned providers. For security reasons, we propose to interconnect the concerned RADIUS servers by using the IPSec protocol [76].

Roaming SIM-IP Card Configuration

We subdivide the SIM-IP connection procedure into three main phases:

1. The card physically connects to the visited network; authenticates itself and the network, negotiates link layer keys and establishes a protected L2-link. This is described in the previous section.
2. The card executes an extended DHCP query and obtains its own IP address and the IP address of the SessionDB server. The card connects to the

SessionDB server using the proprietary SIM-IP Roaming Update Protocol (SIRUP). Using AAA over the local AAA server, the SessionDB can submit session information to the card's home network. During the SIRUP conversation, the card obtains the configuration information, the necessary available service and QoS class descriptors, etc.

3. The card presents a login possibility to the user proposing the collected information on the visited network, in particular the available services, service classes and prices. User authentication is particularly easy since it remains terminal-internal between the OS and the card. The user can log in using her user/password combination, PIN, etc. From this point on, the user session starts. This can be charged by issuing the necessary command over SIRUP to the SessionDB. The latter can gather accounting data from the visited SPN and submit to the AAA server that can forward it to the home network.
4. Depending on the user's choice, the user's profile authorizations and/or required service (launched applications, explicit choice, etc.), the card negotiates and reserves the QoS for the network service in the SPN. The card can store the accounting summaries and submit accounting requests.

The SIM-IP Roaming Update Protocol (SIRUP) is one of the key elements of this process. This protocol is to be developed in detail but since the card is already equipped with a TLS protocol stack, the used cryptographic functions (RSA, SHA-1, etc.) and HTTP, it can be based on HTTPS. All smartcards have hardware support for RSA. The private and public key operations generally necessary for e.g. TLS HANDSHAKE protocol [77] can thus be efficiently supported. Some cards also have cryptoprocessors supporting symmetric encryption functions (e.g. the JavaCard API specification includes 3DES support [138]). This permits the usage of TLS encryption (TLS RECORD protocol [77]).

The usage of HTTPS is particularly convenient since there are provisions for applet handling and Java integration in HTTP. Some additional optimizations are possible: e.g. since the AAA server and the EAP-TLS method have already negotiated a TLS master secret, we can directly proceed with the next TLS phase that encapsulates the HTTP transfers (leveraging the session resume feature). That requires TLS master key delivery from the AAA server to SessionDB after a successful card connection (this optimization is generally not possible in case of roaming and the whole TLS has to be used in that case).

During the SIRUP TLS conversation, the SessionDB installs a bidirectional IP-to-user mapping which it extracts out of arriving TLS-protected IP packets. It gives access to this information to all registered (i.e. trusted) SPN services.

Additionally to the information on the available QoS classes and user services, the card could download and install new applets. Those could be proxies or, in the home network, core service updates (security updates, etc). The card uses the obtained network information to properly configure its service access points (SoC proxies or core services).

Alternatively to proprietary SIRUP, providers could use some extensible service discovery protocols e.g. UPnP which has already been presented in Section VI.3.3.

Roaming Service Access

The L2 security measures (e.g. encryption) spawn from the terminal equipment to the network AP. Hence, after the packets finally arrive in the IP-based part of the provider network, no card/user identity information remains included in the latter. So, how could we possibly identify the user (or the card acting on user's behalf) to provide a

personalized service access? The only identity information still included in the IP header is the source IP address. Due to the simplicity of the so-called *IP-spoofing attack*, this mechanism usually can not guarantee a reliable identification. We believe that with some changes this simple but powerful identification method can be used in a relatively secure way.

The IP-configuration information provided to the card by DHCP is transported over the per-card protected logical link, so it can not be sniffed or altered by attackers. The host OS reacts with an own DHCP request as soon as the card indicates a successful link establishment. This message can be intercepted and replied to by the SIM-IP card, thus providing the OS with the necessary IP information. From now on, every packet issued by the OS can be verified by the SIM-IP card for the validity of the source IP address. With the assumed security of the SPN, the obligatory L2 protection between the card and the AP and the impossible IP-spoofing by the authorized users (due to card-based control), no packets with incorrect source IP address can enter the network.

Servers are to be located in the IP-based part of the SPN that is physically or logically (subnetting, firewall or packet filter) protected from whichever access originating from the public networks like the Internet.

The SessionDB provides access to the IP-address/user mappings to the trusted IP-based services. It installs appropriate routing rules for the card's source IP e.g. allowing or disallowing the Internet traffic or access to a particular service. In parallel, when a source IP requests access to a service, the service control entity (CONTROL in Figure VI-12) can identify the corresponding user by interrogating the SessionDB with the source IP address as key.

To support secure roaming, the two concerned providers maintain an IPSec tunnel. The traffic from the card's source IP to its home network should be routed over this IPSec-tunnel. Then, the IP-to-user mapping obtained from the visited SessionDB can be used by the home SPN to admit and charge the service access. On the other hand, this access can also be handled by some proprietary method defined by the provider and implemented in the card (e.g. a VPN module, for which the card can negotiate a security association).

VI.4.5. Possible Terminal Implementation

Unlike GSM or UMTS, most wireless technologies do not have provisions for smartcard support. It seems unrealistic to demand that e.g. 802.11 network adapters require a smartcard for their activation. Thus, we presume that users can principally establish connections over the available network adapters without SIM-IP. Besides, some providers may decide to propose basic services - i.e. hotspot alike - to completely unknown users. Users can use this service without the SIM-IP card directly over the available interface. In this case, they appear as guests.

Using SIM-IP, users obtain a managed virtual computing environment with QoS guarantees. In this case, the question comes up how SIM-IP can control the user network traffic. The answer to this apparent problem is based on the trust between providers and SIM-IP.

We propose to implement the SIM-IP card support in form of a virtual network adapter concept as the TUN/TAP [139] virtual interface concept in such operating systems as Linux or BSD. This concept is often used in conjunction with the integrated VPN support in modern OS. A software brick (VPN module) implements the VPN protocol (typically consisting of a data transport and a session signaling parts). It then uses an available physical interface with the associated IP stack to establish connections with a VPN

gateway, e.g. to a trusted Intranet. Using the OS API, it then creates a virtual interface pointing to itself. The data from this interface are received by the software module, encapsulated, protected and sent out over the physical interface according to the internal states established by the session signaling part. For applications, the connections over this new (virtual) interface appear as connections within the home network. An example for such systems are PPTP or L2TP-based user VPN drivers.

With SIM-IP, the same mechanism can be used. Herein, SIM-IP concentrates on the session signaling part. The data are transported by the physical interface using the L2 protection measures, without any higher-level overhead on the precious wireless link (as with VPNs). The necessary key material is delivered exclusively by the SIM-IP. This is comparable to the GSM SIM concept: GSM SIM is only responsible for the authentication (A3) and key derivation (A8); the actual encryption (A5) is implemented by the handset.

SIM-IP is the only entity in the terminal capable of such session key derivation. Given that SIM-IP itself is considered secure and tamper-resistant, the user can not obtain an uncontrolled network access. Hence, the first important insight is: even if a user tricks the internal SIM-IP controls, the abuse discovered in the SPN will be correctly attributed to the user's account.

A valid user can try to bypass the SIM-IP once it has established the link. Hence, the connection from SIM-IP to the network adapter will be suppressed and the virtual interface mapped directly to the physical interface. Given the probable shape of the 4G terminals, this attack would need major competence in embedded system design. The important point however is that without the SIM-IP control the L2 session over the physical interface will expire very soon. The session will be shut down by the access point. Besides, all sent data can be counted at the AP and correctly charged to the user account.

A more intelligent attack would be to let the SIM-IP card attached and to inject packets directly to the physical interface. That would allow for e.g. uncontrolled traffic bursts over the wireless link. For that reason, SIM-IP steadily counts the packets coming from the OS before it forwards them over the network adapter. Taking into account transmission errors, it can now keep an important snapshot of the amount of data sent out in a conforming way (i.e. through SIM-IP). SIM-IP can use SIRUP to fetch network's snapshot from the SessionDB. In case of discrepancies between the internally obtained and the network snapshots, reports can be generated to the home provider. Furthermore, in this case the user session can be shut down by the SIM-IP.

Because of the missing physical contact between the SIM-IP and the network interface, abuse by valid users is possible even with the SIM-IP card in use. We conclude that in spite of the presence of SIM-IP, additional checks of contract compliance within the SPN remain indispensable. Alone, the SIM-IP card can motivate (attack complexity) and promote (additional comfort) such compliance but neither enforce nor guarantee it.

VI.4.6. Performance Expectations

Our approach does not require any changes to the used standard mechanisms. SessionDB is primarily a slightly adapted HTTP server with an associated database. The recent advances in Internet technology prove that very high loads can be supported by this technology.

We address the performance of the new introduced entity, the SIM-IP card. We then make an estimation of the delays associated to the proposed access procedures.

SIM IP Card Performance

Smartcard is a small piece of silicon, whose area is about 25 mm², that embeds a CPU, and a non volatile memory. Maximum memory size of existing components is around one megabyte, but chips including two megabyte of FLASH memory are already available. CPU are designed with a 8, 16 or 32 bits core; their computing capacities range between 10 and 100 MIPS. However, it should be noted that the security management provided by the smartcard operating system significantly reduces the processor time allocated to the embedded applications.

TCP/IP stacks have been recently embedded by various smartcard manufacturers (see for example [136]). The maximum data throughout supported by the ISO 7816 standard is about 200,000 bps, but many proprietary physical interfaces are proposed by founders, with a few Mbps.

According to the elements mentioned above, smartcards could be seen as a DPX266 PC, that was used ten years ago, with 10 Mbytes of memory, a 100 MIPS CPU and a 10 Mbps network interface.

Because most smartcards include a dedicated cryptographic processor, they are able to execute complex RSA calculations (2048 bits key size) in less than 100 ms. As an illustration it has been recently demonstrated that commercial Javacards are able to process, with an acceptable performance, the TLS protocol, and therefore are suitable for authentication in wireless LANs [140].

4G research has just started and first working system prototypes are not expected before 2009-2010. Given the amazing recent advances in the smartcard development and anticipating further progress in the coming 5-6 years, it seems reasonable to use smartcards more extensively in the next generation wireless networks.

Network Access Performance

The network access performance is not critical to the system performance since the four described phases occur only on network login access (i.e. relatively rarely). The first phase is a standard 802.11+802.1X access phase. Independent measurements show that the overall practical delay of this access procedure is about 1s [117]. In roaming cases this can be slightly higher as confirmed by our own experiments.

The card configuration phase implies a combined DHCP and SessionDB interrogation. DHCP delays are reported to be about 1.7-1.8s [141]. Then the card fetches the XML service description file via HTTPS-based SIRUP. This connection remains local to the visited network. HTTPS connections to web servers using old hardware, PDAs and general cryptolatency studies are evaluated in [142][143][144]. Our own practical experiments with the `openssl` toolkit[1] emulating respective operations in a local network show that this latency always remains under 500ms (average time was less than 200ms). We use the session-resume feature to accelerate the TLS handshake (2 RTTs less) and to avoid unnecessary private key calculations in the SIM-IP card.

The overall estimated delay of both initial phases is thus about 3s. This has to be repeated for every policy- and profile-allowed network on every interface. Theoretically, the overall discovery delay could reach high value. However, practically we expect the number of usable access networks to be less than 10. Failures (802.1X authentication failure, no SessionDB, etc.) result in break-up and thus accelerate the discovery. Finally,

[1] The OpenSSL Project (http://www.openssl.org) implements a simple HTTPS server and client for demonstration purposes.

some procedures can be run in parallel given that a terminal is expected to have about 4-5 network interfaces. Nevertheless, there clearly is room for further protocol optimizations.

The third phase is interactive and can not be quantified.

The fourth phase represents a standard COPS procedure. COPS results are taken from [145] and [146].

Table VI-5 Transmission and treatment delays for typical COPS messages

	Transmission and treatment times of COPS messages in ms		
	average	Min	Max
Message REQ (PEP→PDP)	1.33	-	-
Message DEC (PEP←PDP)	37.83	-	-
Message RPT (PEP→PDP)	38.76	-	-
Total	77.92	76.5	78.3

VI.4.7. Conclusion

In MMQOS, we have designed an open and flexible architecture supporting a dynamic adaptability to overcome the dynamicity of the encountered environment. This adaptability, the dynamic service reconfiguration and the adjustment of link layer security and QoS parameters are carried out transparently by a newly introduced SIM-IP card module in the user equipment. The SIM-IP card thus virtualizes the discovered environment by defining a new, complete user interface and by hiding the actual implementation in an encountered network from the user. In this sense, the SIM-IP module goes beyond both the security mechanism abstraction of the GSM/UMTS networks and the transparent home network service access provision of the popular VPN technology.

From the user's point of view, our approach guarantees a stable service environment. From the provider's point of view, our system uses the next available resources and thus better fits the service environment. It optimizes the overall service access in the system by avoiding unnecessary routing to the home network. The necessary card roaming, service discovery and access mechanisms are part of the SIM-IP implementation. To some extent this allows system-compliant proprietary solutions per provider resulting in more flexibility for providers.

Besides, user mobility issues were studied in the MMQOS project. We believe that link layer mobility support should be used to its full extent whenever possible. For global terminal mobility our solution already provides full and transparent roaming support. For session mobility and service continuity, we believe that a SIP-based approach is a good candidate. SIP/COPS interaction studies both for IntServ and DiffServ architectures have been carried out in [145].

Principally, our solution seems suitable as it can qualitatively solve the multi-layer access problem in future 4G networks. Further protocol and mechanism optimizations are possible. These are necessary to solve the problem quantitatively, providing full user-friendliness. We are convinced that the ongoing progress in smartcard and 4G development could result in using our approach as a reference architecture.

VI.5. Comparative Evaluation of the Proposed Architectures

VI.5.1. Comparison of RESACO and MMQOS Approaches

From our point of view, both RESACO and MMQOS projects address the same issue of the dynamic adaptation to the service environment in the 4G scope.

RESACO addresses this issue by defining an adaptable network. RESACO studies the aspects of the network plane adaptation to the user's profile at network access time and to the user's request during the session. In particular, we have studied and implemented the network edge equipment capable of such dynamic adaptation. Using its adaptable components and different adaptation mechanisms, this approach transforms a real SPN into a virtual environment which the user is accustomed to.

MMQOS addresses the same issue by using a different approach. It adds a new trusted abstraction element within the terminal: the TCP/IP-capable smartcard. This smartcard hides the particular implementation details of the different SPNs and provides users with the same, accustomed environment.

In both cases, a minimal common architecture of any SPN is necessary. In RESACO, this applies to any procedure on the user link (e.g. to the service discovery procedure and the corresponding architecture within the SPN control plane). In MMQOS this applies to the procedures between the card and the SPN.

Table VI-6 Comparison of different related approaches to RESACO and MMQOS

Criteria	CHOICE	GUIDE	PANA	RESACO	MMQOS
Access Control	L3	L3	L7	L2 native	L2 native
User security functions	Web-based auth, proprietary authorization and tag-based last-hop QoS	User auth, L3 data last hop encryption	Strong auth., full AAA support	Strong L2 auth, full AAA, full L2 security, system security	
Data packet overhead	L3 tag	L3 tag	No	No	No
Client OS modifications	special packet tagging module	packet tagging in a modified MIPv6 tag	PaC installation	No (only user-space applications used)	virtual interface driver using SIM-IP card
Environment stability	No	No	No	Yes	Yes
Supported L2 technologies	Agnostic	Agnostic	Agnostic	Various	Various
Mobility support possible	Nomadicity	MIPv6 terminal mobility	Nomadicity	L2 mobility, global roaming	L2 mobility, global roaming, SIP mobility
Network management	proprietary or undefined	by AAA	out of scope	Provider intern, can use COPS or SNMP	Policy based using COPS

VI.5.2. Comparison to Previous Work

Since our approaches explicitly address 802.11 WLANs, we can compare these to different previously published approaches mentioned in the related work section (see IV.3.3). In Table VI-6, we qualitatively compare two different PAWNs and IETF's PANA (generally representing here higher-level access control approaches) to RESACO and MMQOS proposals described above.

VI.6. Decentralization of the Control Plane

After having treated different issues on the user-network link, in the AN and in the network plane of the SPN, in this section we concentrate on the SPN control plane. We first discuss the control plane organization in light of the existing protocols and the architectures presented earlier in this work. We discuss the shortcomings of the existing solutions taking into account the diversity of SPNs in the 4G scope.

VI.6.1. Problem Statement

Logically, the control or management plane is the intermediate plane between the administrator and the equipment. It simplifies the network administration by taking control of all standard tasks and by making the decisions in all standard situations as previously defined by the administration by means of rules and policies. Since there is logically one administration but a variety of equipment to be managed, the control plane necessarily provides some sort of n:1 relationship.

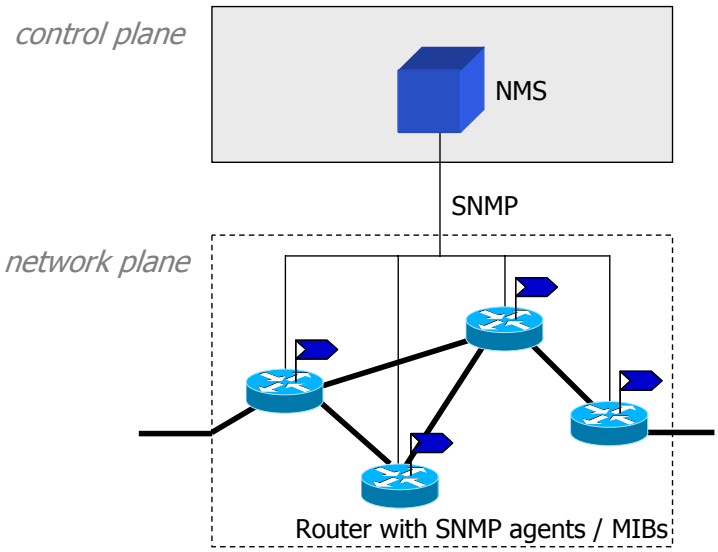


Figure VI-13 A typical SNMP-based network management architecture

Technically, the control plane typically implements this by centralizing the control of the dispersed equipment on a single control point as illustrated at the example of an SNMP architecture [84] in Figure VI-13. SNMP implements the agent/manager paradigm. Here, the central Network Management Station (NMS) in the control plane collects management data from the agents installed in the network plane equipment[1]. The NMS

[1] Mapping this to our nomenclature, the NMS is the central PDP and the agents are PEPs

then serves as the central control point for the administration accessed from the business plane. Another example is AAA framework [90], classically implementing the client/server paradigm[1]. Here the network access servers (NAS) send requests to a central authentication server (see Figure VI-14). This concept is widely used today in different systems. A prominent example is GSM with its hierarchical architecture, central HLR, AuC, etc.

Indeed, the technical centralization of the control plane is not always optimal as they bear different scalability, robustness and cost issues. These are due to the underlying concept described above – they are mostly independent of a particular framework. For instance, in our architectures RESACO and MMQOS presented in Sections VI.3 and VI.4 respectively, these issues apply to the AAA protocols like RADIUS [91] and DIAMETER [93] in the same manner as to the network management protocols like COPS [85] and SNMP [84]. The exact impact of the issues might change depending on the protocol, the SPN and the AN, however principally they exist in either case.

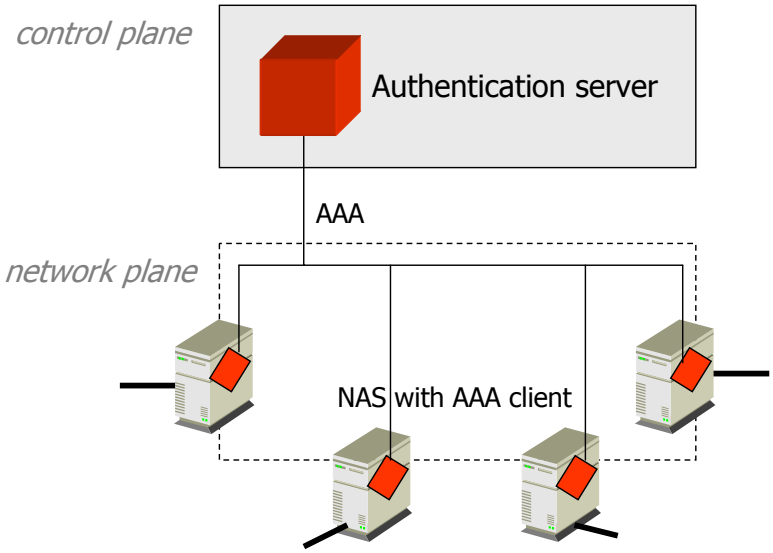


Figure VI-14 AAA model with Network Access Servers and a central AS

Because control planes principally target provider needs, the control planes of RESACO and MMQOS are inspired by GSM. Consequently, these control planes make use of central control and decision points. However, we clearly decouple the internal control plane organization from the user-network and provider-provider interfaces. Thus, each provider can follow a different concept for the fully internal control plane organization.

To achieve a broader coverage and AN availability in the scope of future generation networks, it seems essential to support not only different technologies but also different businesses managing the deployed SPN/ANs. Thus, the proposed SPN architectures, and in particular the SPN control plane, should allow different SPN instantiations from national scope service providers to very local providers. From the technical point of view, providers should be able to choose their operation mode, to expand and to shrink when necessary, etc. without having to replace the used control plane or the existing network plane equipment.

[1] New AAA protocols like e.g. DIAMETER actually implement a mixed client/server agent/manager paradigms.

In this section, we present an integrated access control and management system that implements a fully distributed control plane still enabling an easy centralized control by the administration from the business plane. To provide a concrete example, we base our discussion on a 802.11 WLAN using 802.1X access control and develop COMPASS (**CO**nfiguration **M**anagement **P**2P **A**ccess **S**ecurity **S**ystem)[1].

In the following sections, we first analyze the issues of centralized control planes at the example of the AAA framework for both big installations and small local systems. We then formulate the requirements on the distributed access control architectures. We finally present the main idea and some implementation details of COMPASS.

VI.6.2. Limits of the AAA Framework

Scalability

In the classical AAA framework, a central authentication server (AS) is used by network access devices (access points in our nomenclature) to verify credentials presented on the user link. Depending on this verification and local policies, the AS replies with an access accept or reject message[2]. Thus, the AAA framework is mainly used as an internal interface to the user management.

In our model, AAA is expected to be used as a universal backend access control for different services (effectively implementing *single sign-on*). Typical access points are e.g. network edge devices and service authorization verification. Here we concentrate on network access control using 802.11 with 802.1X as an example.

Although the user data traffic never passes through the AAA server, the latter is the only decision point for session admittance and thus represents a *single point of failure* (SPF). If the central AAA server fails, no access point administratively submitted to this server can accept new sessions. This is illustrated in Figure VI-15.A. The existing sessions on such access points will have to be shut down at the next re-authentication. Generally, possible reasons for a server failure could be some of the following:

- AAA server overload
- Host system failure
- Partial network failure
- Partial power down

Among these possible causes, an AAA server overload can occur even under normal operating conditions. The exact saturation point of an AAA server is influenced by various parameters such as:

- Number and type of services used against the AAA server
- Overall user number
- Per-service properties
- Defined user session duration
- User authentication method

[1] While a *radius* connects a point on the circumference to its center, a *compass* is the direct interconnection of all circumference points.

[2] In the subsequent development, this has been extended to a model in which the user equipment opens a logical communication channel to the authentication server over the access point. Although technically identical, this latter user to server communication (implemented e.g. by EAP usage [94]) can require per-session state at the AS and thus principally increases the complexity of the server logics.

- Use of accounting
- User mobility and its influence on this particular service

Depending on the used authentication method, the server processing of the authentication data could vary from a stateless answer (as with PAP [91]) to stateful per-session, two-phase data processing with multiple exchanges, private key operations, etc. (e.g. EAP-PEAP with MS-CHAPv2). Since e.g. the new security for 802.11 depends on the usage of mutual strong authentication methods with key derivation [81], the load per user is typically higher than with classical dial-up access technologies.

To provide a more precise and robust snapshot of the active network sessions, modern AAA protocols provide support for periodic accounting (*interim accounting*). The exact period typically depends on the defined session duration and, multiplying with the number of access points, massively increases the number of messages to be processed at the server.

As always, scalability problems can be mitigated by installing additional AAA servers. That could be done e.g. by installing an AAA server per AN or a given AN could be further subdivided in AAA-governed subnets of an appropriate size (which depends on the criteria cited above). This is shown in Figure VI-15.B where the AP at the left authenticates over the AAA Server 1 while the remaining APs authenticate over the AAA Server N. Modern AAA protocols (RADIUS [91], DIAMETER [93]) provide measures for AAA server interconnection (so-called *proxying*) achieving such subdivision without user mobility limitations: a user whose profile is managed by the AAA Server 1 can still access the service over all connected APs.

However, such multi-server solution comes at the price of additional, costly server installations, their maintenance and an increasingly complicated SPN infrastructure with independent local user databases, interconnected AAA servers in different ANs, etc. A continuous network growth (due to e.g. enterprise expansion) is difficult to follow with this method since a new AAA server is difficult to add due to the new database repartition and necessary trust relationships. If the user databases are completely replicated, consistency mechanisms have to be used to enforce the same content in all databases. This is difficult because changes can occur in different databases at the same time. If the database is not replicated to every AAA server, then each AAA server becomes a SPF for all users managed in its database. Obviously, there is an undesirable trade-off between the control plane performance on one hand and its complexity on the other.

The installation of multiple AAA servers authenticating towards a common user database as shown in Figure VI-15.C does not mitigate the scaling problem. It principally equals the first case (Figure VI-15.A) with scalability problems shifted to the database. Extra cost must be taken into account for the server software, installations and its maintenance.

Small Networks

The centralized concepts like a central AAA server are not well suited for very small installations. The main problem is the cost of a reliable central installation and the cost of its operation. We principally would like to develop a control plane that can be similarly used in private networks, enterprise deployments and small provider SPNs. Due to its versatility, the management of an AAA infrastructure generally requires profound network knowledge and competent administration. The administration effort and the additional equipment, software and maintenance cost are difficult to amortize in a very small installation.

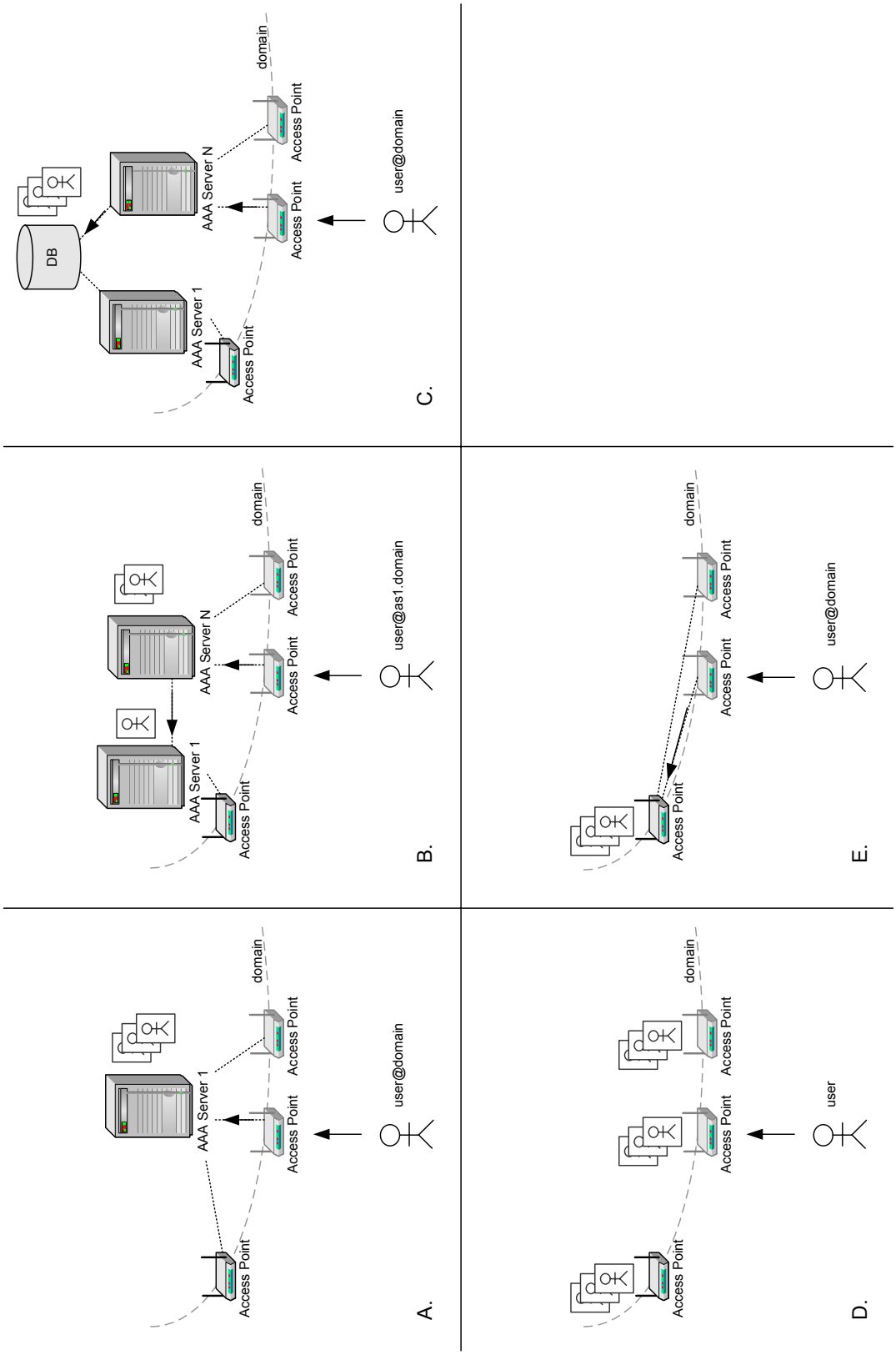


Figure VI-15 Possible AAA-based access control architectures

For instance, that makes it difficult, especially for small enterprises, to use the currently available access control solutions for WLANs: these are either not secure enough or their

security is not affordable. For this reason, IEEE 802.11i provides a pre-shared key (PSK) mode for standalone APs. However, in this mode it is almost impossible to provide access to occasional visitors and to different user groups. With current implementations, the mode with the AAA server features a more extensive access control, in particular allowing a much finer user authorization, per-user secrets and accounting support. Furthermore, if a WLAN installation based on PSK has to be extended to several APs, the extension principally comes at the cost of reduced security (all APs have to be configured with the same, network-wide shared secret) or it results in a mobility-limiting, predefined user-to-AP assignment as shown in Figure VI-15.D. This concrete example easily extends to arbitrary services, underlining the current problem: breaking the central control currently forces administrators to base the system security on system-wide secrets.

The only existing alternative is to make all new APs authenticate users over the first AP acting as a local AAA server and containing user profiles. This is shown in Figure VI-15.E. However, though practically easier to achieve in a small network, this solution comes down to the installation of a central AAA server, with particularly limited resources. This solution can not be easily extended. Also, the robustness of this mode is an issue since this special access point represents a SPF. Furthermore, the existing integrated AAA servers are intentionally kept relatively simple and typically do not provide the whole functionality of a dedicated AAA server. Also, an isolated AP is hardly capable of storing the accounting data due to storage limitations.

VI.6.3. Requirements

The limits of the AAA framework motivate the search for access control frameworks without additional central elements. This is especially true for low cost environments like 802.11 WLANs. For the latter, we namely formulate the following requirements:

- *Scalability*: the proposed solution should be able to seamlessly follow the network growth.
- *Fault tolerance*: no SPF is allowed. Additionally, the impact of a partial system failure should be limited to the failed components in an average case.
- *Network extensibility*: a new access point should be easy to add to the access control infrastructure (not more complicated than an AAA client configuration).
- *Compatibility*: no changes on the user link implied by the changed control plane.
- *User management*: the proposed solution should provide an AAA-like user profile support, with the same or equivalent possibilities of user management. User mobility constraints must not be implied by the architecture, i.e. every user must be potentially able to access every part of the WLAN. The solution should be capable of providing accounting support.
- *Network management*: the proposed solution should provide a possibility to manage the connected APs (e.g. check AP status, AP load, APs online, failed APs, etc.), to update their system configurations or system firmware.
- *No negative impact on normal WLAN performance*: a proposed system should not reduce network performance on the user link.

As can be seen in the classical AAA model (see Figure VI-14), the access points integrate a part of AAA functionality, the AAA client. Our idea is to replace this part by another functionality without changing the rest of the access point functions. Instead of the client/server paradigm used with AAA, we use new telecommunications paradigms.

VI.6.4. New Paradigms and Concepts

Main Idea

To meet the requirements defined above we propose to integrate standard network access mechanisms on the user link with an architecture leveraging recent advances in the wired peer-to-peer (P2P) networking in the control plane.

To reduce costs and to achieve natural scaling capabilities, we propose to use the access points directly for the storage of data relevant to access control and network management. However, since the resources of each access point (AP) are limited, the idea is to distribute the administrative load over all access points. Thus, every AP holds only a part of the whole management database. To achieve that, the AAA-client part in the access point is replaced by a P2P part.

The difficult part is to provide a scalable and robust mechanism for distributed data retrieval. Herein, the problem is not the data transfer but locating the data [147]. First, to fulfill the scalability and fault tolerance requirements, no entity is allowed to have global network knowledge. For instance, no single AP is allowed to have an index of all data records in the overlay. Second, network wide broadcasting (e.g. "who-has data-record X?") is not allowed for efficiency and scalability reasons. Finally, in the given environment we can not accept threshold-based, iterated limited-scope broadcasting since the request iteration can result in randomly increased search delays, timeouts and, generally, in decreased service quality. Moreover, this mechanism does not always find the available data (namely, when the data are held by a node located more than threshold hops away).

Distributed Hash Tables (DHTs) have been designed to overcome these difficulties. We want to use DHTs to store and retrieve AAA database data distributed over the APs.

Distributed Hash Tables

A distributed hash table (DHT) is a hash table divided into multiple parts. These parts are distributed to certain clients now typically forming an overlay network. Such a network allows the user to store and retrieve information in *(key, data)* pairs as known from traditional hash tables. They need specific rules and algorithms for the distributed access. Famous examples of distributed hash tables are file-sharing/P2P networks such as EDonkey2000 or Kazaa. Each node taking part in such a P2P network is responsible for one part of the hash table called *zone*. By this means we avoid a central network entity managing the complete hash table or its index.

Every node participating in such a P2P network manages its part of the hash table and implements the primitives: `lookup(k)`, `store(k, d)` and `delete(k)`. With `lookup(k)` a node searches the P2P network for a certain hash key `k` and receives the data `d` associated to key `k`. As every node possesses only a fraction of the complete hash table, it is possible that `k` is not in the node's own fraction. Therefore, every DHT defines an algorithm to find that particular node `n` responsible for `k`. As this is done on a hop-by-hop basis with each hop "nearer" to `n`, it is called the *DHT routing algorithm*.

The primitive `store (k, d)` stores a tuple consisting of a key `k` and the associated data value `d` into the network, i.e. `(k, d)` is transmitted to a node responsible for `k` using the same routing technique as with `lookup`. With `delete(k)` an entry is removed from the hash table i.e. the node responsible for `k` removes `(k, d)`.

So, P2P based overlay networks utilize their own routing or data forwarding mechanisms [147]. These are optimized so that each node has only a very local view of its network

neighborhood. This property is necessary for a good scaling since the per-node state does not necessarily increase with the network growth. Routing is deterministic and there are upper bounds for the number of hops a request has to pass. Most P2P networks feature logarithmic behavior with a total number of nodes.

Several DHT-based P2P systems have been proposed so far but not all of them suit the needs of our application. We have to choose a P2P system tailored to the restricted resources of an ordinary AP. We namely compare three DHT-based P2P systems CAN [147], Chord [148] and Pastry [149].

Without going into structural or design details of these networks, we present CAN and then give a short comparison of DHT properties that seem crucial for our work.

Content Addressable Network (CAN)

CAN [147] is a highly scalable DHT based overlay concept. Defining a standard hash table user interface (as discussed above), CAN provides the following mechanisms mainly concentrating on scalability and robustness:

- overlay construction (node join/node bootstrapping)
- node leave
- routing algorithm

CAN's hashtable index is a d-dimensional Cartesian coordinate space on a d-torus. Each node is responsible for a part of the entire coordinate space. Figure VI-16 shows an example of a 2-dimensional CAN with 5 nodes (A, B, C, D and E) [147].

CAN's mechanisms are optimized to minimize the per-node state. In CAN every node holds the zone database corresponding to the assigned coordinate space and an overlay neighbor table. The size of this latter depends solely on the dimension `d`. The standard mechanism for zone assigning achieves a uniform distribution of the index over the nodes.

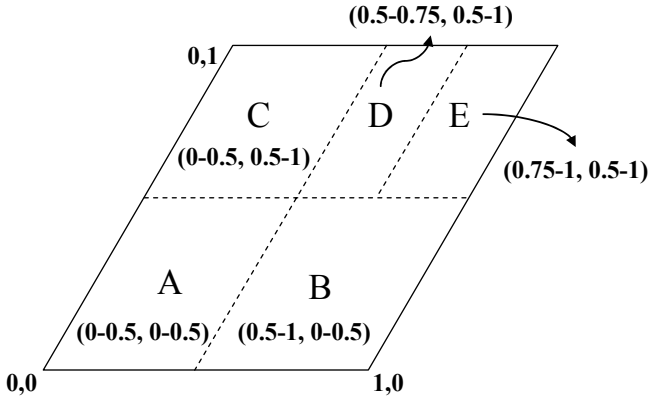


Figure VI-16 A 2-dimensional CAN

Per default, CAN makes use of an overlay construction procedure (called *bootstrapping*) based on a well-known DNS address. This enables every joining node to obtain an address of one or more CAN bootstrap nodes. When receiving a request from a new node, a bootstrap node simply answers with IP addresses of several randomly chosen nodes that are in the overlay. The join request is then sent to one of these nodes. The new node then randomly chooses an index address and sends a join request for this address to one of the received IPs. CAN uses its routing algorithm to route this request to the node responsible

for the zone where this address lies in. The solicited node now splits its zone in two halves and hands out one of the halves, the related zone database and the derived neighbor list to the joining node.

For example, the CAN in Figure VI-16 is a possible result of the following scenario:

- A is the first node and holds the whole database
- B joins and gets a half of A's zone split along the x axis
- C joins and randomly gets a half of A's zone split along the y axis
- D joins and randomly gets a half of B's zone split along y
- E joins and randomly gets a half of D's zone split along x

The routing in CAN is based on greedy forwarding. Every request contains the destination point in the index space. Every receiving node not responsible for the destination point forwards this request to one of its neighbors whose coordinates are closer to the point than its own.

To improve the performance (eliminate latencies, achieve better robustness, etc.), CAN features different parameters:

- *Adjust the dimension* `d`*:* the number of possible paths increases with dimension thus achieving better protection against node failures. The overall path length decreases with `d`.
- *Number of independent realities* `r`*:* by using multiple `r` independent CAN indexes within one CAN, r nodes are responsible for the same zone. The overall path length decreases with `r` (since the routing can take place in all realities in parallel and be abandoned on success). The number of really available paths increases. The data availability increases since the database is replicated `r` times.
- *Use of different metrics, reflect the topology in the overlay:* CAN can use a different routing metric. The underlying topology can be reflected in the overlay.
- *Node peering:* the same zone can be assigned to a group of nodes thus decreasing the overall number of zones and the path length.
- *Using multiple hash functions:* this is comparable to multiple realities since each hash function constructs a parallel index entry.
- *Caching and replication of data pairs:* "popular" pairs can be cached by nodes and thus replicated within the database.

A complete comparison of the parameters and their effect can be found in [147].

Comparison of Different DHTs

In Table VI-7, "#Hops for lookup/store" is the expected number of different overlay nodes a request for a lookup or store has to pass. The second criterion is the number of elements each node has to store in its neighbor or routing table. Both properties are expected values for a well-balanced DHT overlay network.

Table VI-7 Properties of common DHTs (n total number of nodes in the network, d is a fixed small number parameter to CAN)

Property	CAN	Chord	Pastry
#Hops for lookup/store	$O(dn^{1/d})$	$O(\ln n)$	$O(\log_{2^b} n)$
#Elements in routing table	$2d$	$\ln n$	$b*\ln n$
Used in	Secure Service Directory	CFS file system	Oceanstore, Scribe

If the total number of hops is known in advance one can adjust CAN's dimension `d` to:

$$d = ln\ n$$

to get a logarithmic behavior for the number of hops in a lookup/store request in CAN also. In comparison to CAN, Chord and Pastry seem to perform better regarding the number of hops involved in queries. This means less communication overhead and shorter delays. On the other hand, CAN's big advantage is a constant *O(1)* restricted memory consumption which is known at node setup time, prior to network use. As a result we choose CAN because of the constant number of elements in its routing table. On a resource limited access point, main memory is much more crucial than the expected better behavior regarding communication overhead for lookup and store requests.

VI.6.5. COMPASS: Decentralized Management Architecture for WLANs

In the following sections, we present COMPASS, our decentralized management architecture for 802.11 WLANs. COMPASS is an instantiation of our ideas discussed above, showing how decentralized access control and management can be integrated with an existing popular access technology. We first present the basic system architecture introducing the main entities in COMPASS. We then briefly discuss the trust relationships in our system and the resulting possible security functions and mechanisms. Then we show how different system tasks are carried out including necessary preliminary AP configuration, overlay construction and typical user management tasks. Next, we show how users access a network controlled by COMPASS. Finally, we address different issues such as data delivery in the overlay, auxiliary host support and mechanisms to achieve higher fault tolerance. Finally, we discuss the system profile storage capabilities and give a qualitative comparison to the different possible AAA-based architectures discussed above.

Basic System Architecture

Based on standard TCP/IP networking in the core network, our P2P management network is formed by the used 802.11 access points. This is illustrated in Figure VI-17. Every access point acts as a P2P node building a logical overlay network over the physical core network. This overlay stores different logical databases, primarily user and management databases (DB). The user DB stores AAA-like user profiles. The management DB helps the administrator manage all connected APs and stores AP settings expressed in the respective syntax (e.g. 802.11 MIB variables, proprietary manufacturer settings, etc).

On user demand, the solicited node retrieves the correspondent profile by using the overlay's lookup method. Using the retrieved profile, the serving AP follows the usual 802.1X procedure acting as Authenticator with a local Authentication Server [80].

Additionally, we provide an optional possibility to include any number of assisting nodes (so-called *auxiliary nodes*, e.g. network administrator's console) in this P2P network.

Security and Trust in COMPASS

To enable all possible overlay topologies, all nodes participating in the P2P network trust each other to route the requests and to retrieve, store and remove data correctly and securely. The P2P network acts as an entity, accessible over any connected AP.

With n APs and no central entity it is convenient to express this trust relationship by means of public key cryptography using e.g. signed certificates. The signing authority (certification authority, CA) does not form an SPF since it signs the nodes in some preliminary step (initial AP configuration described later) and it is not necessary for

signature verification in the operation mode. This approach enables secure communication establishment between two arbitrary participants at any time with n secrets for n nodes. The defined identity of an AP is its MAC address of its wired interface connected to the CAN.

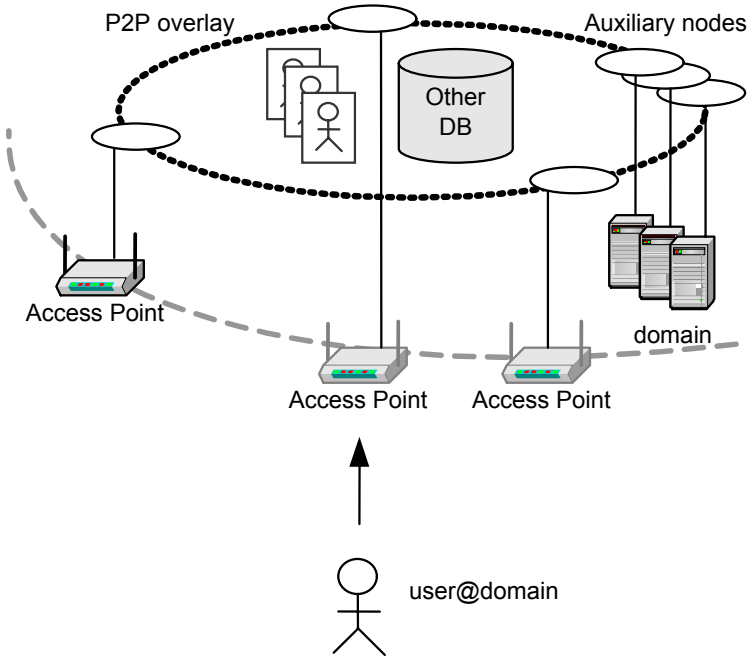


Figure VI-17 Main entities in COMPASS

Depending on the anticipated risks in particular environments, the implemented overlay should provide data transfer, routing and administrative overlay access protection. Suitable mechanisms seem to be TLS, IPSec, etc. The existing trust relationships can be used to achieve this goal.

Preliminary AP Configuration

Each AP needs a minimum configuration prior to its deployment in the network. This is primarily necessary for a secure management access to this AP, the overlay discovery but also for classical 802.11 settings.

The trust relationship with the AP is expressed by the installation of a signed certificate on every AP. Further, the administrator defines a local admin login (user/password pair) and adjusts the usual 802.11 parameters (SSID, authentication mode, used channels and throughputs, etc.) Finally, the administrator supplies the bootstrap-address of the overlay network and deploys the AP installing it in the desired location and connecting it to the network.

Bootstrapping (AP Join)

The original CAN proposal makes use of a bootstrap method guaranteeing a uniform partitioning of the index space over nodes (see VI.6.4). However, CAN's standard approach decouples the overlay from the underlying topology. It means that a physical neighborhood does not result in a CAN neighborhood.

We want to be able to tie the overlay to the network topology. This can potentially shorten the handoff-delay: when a terminal hands off from one AP to a (physical)

neighbor AP, the new AP can rapidly retrieve the profile data from the old AP using the overlay (since the old AP is also the overlay neighbor). The other reason for reflecting the physical topology in the overlay is a transparent load balancing: if an AP suffers a heavy load, the administrator likely installs an additional AP in its neighborhood. If the concerned APs are not CAN neighbors, they only share the 802.11 traffic load. If the APs are CAN neighbors, they also share the administrative load.

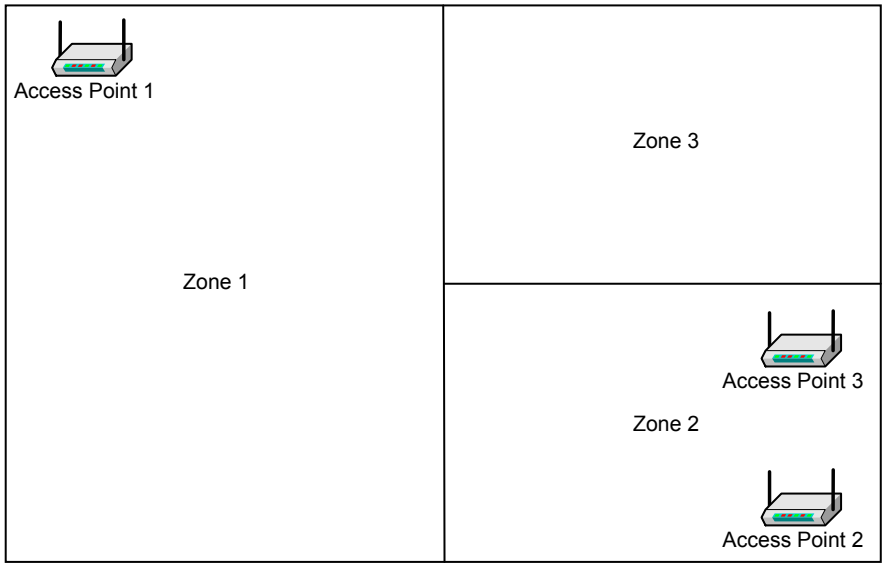


Figure VI-18 Zone management and traffic load

This is represented in the example in Figure VI-18 showing three APs installed in a big hall. At the beginning the initially installed Access Point 1 (AP1) possesses the whole index. With the arrival of Access Point 2, AP1 gives half of its zone to the Access Point 2 (AP2), thus becoming its overlay neighbor (but not necessarily its physical neighbor). Let us now imagine that the user data traffic is particularly high in the lower right corner of the map and relatively low in the upper left corner. The administrator would thus add Access Point 3 (AP3) in the topological neighborhood of the Access Point 2 to process the high wireless traffic load. If we tie the overlay to the network topology, the new AP3 automatically becomes an overlay neighbor of AP2. Thus, it obtains half of the zone database managed by the AP2 (Zone 3). Hence, presuming that the administrator tries to equalize the traffic load, with this approach the zone sizes of the APs decrease in the areas with a high traffic load, thus freeing system capacities for traffic processing. In contrast, the zone of the AP1 remains relatively big – this is however justified by the lower data traffic. Evidently, there is a trade off between the zone management overhead and the WLAN traffic load.

CAN provides an abstract alternative *landmark-ordering* method to reflect the physical network topology to the overlay. However this mechanism is too general and explicitly targets the IP layer topology. In our case, the landmark-ordering mechanism can be defined as following:

- Booting up, a preconfigured AP searches the 802.11 environment for 802.11 neighbor APs. The necessary mechanisms are defined in the 802.11 standard and include an active and passive discovery of neighboring APs of the same SSID [14]. These mechanisms are already implemented by some existing 802.11 chipsets. Using these mechanisms, the joining AP retrieves the (wireless) MAC addresses of all neighboring APs configured with the same SSID.

- The joining AP now sends a discovery request to the predefined DNS address of the overlay (e.g. resolving in one of the overlay nodes in a round robin manner). The discovery request contains the list of neighbor APs discovered in the first step. If the received MAC address list is not empty, the resolved bootstrapping node chooses the neighbor whose zone is the biggest. It is essential to provide a mechanism that allows resolving a wireless link MAC address into the management link IP address without global network knowledge. We achieve this by storing these data into the overlay itself, but other mechanisms could be applicable as well. Thus, the solicited overlay node executes a lookup in the overlay for every received MAC address as a key. The received value is a pair `(IP_address, zonesize)`. The solicited node chooses the pair with the biggest zone size. If the received wireless MAC address list is empty (i.e. the new node does not see any neighbors), the bootstrapping node proceeds as in CAN, i.e. it randomly chooses a point in the index space and determines the node responsible for the zone where this point lies in. In either case it replies with the IP address of the chosen node.
- The join procedure itself is like in CAN. The joining AP now sends the join request to the received IP address. The solicited node splits its zone into two halves and hands out the zone and the associated zone database to the new node. The new node becomes a member of the overlay. It then executes a store command in the overlay, posting its own wireless MAC address, the management IP address and the zone size:

  `store(wless_MAC_addr, data),`

  where `data = (IP_address, zonesize).`
- Each AP is responsible for the validity of that entry.

Following this scheme, the new installed AP automatically becomes an overlay neighbor of one of its 802.11 neighbor AP. The advantages of this scheme are an equal pre-configuration of all APs and the requested loose binding of the overlay to the physical topology. Moreover, this scheme currently avoids any AP to AP communications over the serving interface (i.e. wireless link) since it would need a preliminary new node's 802.1X authentication as a station. (This would be an interesting point to study later, especially given the possibilities provided by EAP-SIG, see Section V.5).

Our method does not affect the scalability since the preconfigured overlay address can correspond to a number of APs. This can be achieved with a round robin DNS address or with a multicast address (e.g. "all overlay APs"). Moreover, the join events (new AP installation, AP reboot, etc.) are expected to be rare – much rarer than the operational procedures like user access.

During an initial deployment (or after a compete system failure), no overlay node is available under the overlay address. This case is considered rarer than join events and requires a special treatment. It can be resolved by weakening the equality of the AP configuration (choose some fixed overlay nodes for join requests) or by dynamically updating the round robin DNS on new node joins and departures (e.g. by using dynamic DNS). Such mechanisms are however out of the scope of this document.

AP Leave

A member AP can leave the overlay as the result of a failure (e.g. AP power down) or because of a scheduled shutdown.

If an AP is shut down correctly, it hands out its zone database to an overlay neighbor with the smallest zone database. The neighbor thus becomes responsible for a second zone. This behavior is inherited from the CAN proposal.

In the case of a sudden failure, the zone databases held by this AP are unavailable during the failure duration and can be completely lost in the worst case. CAN redundancy mechanisms have to be used in that case.

User Add/ User Delete

To add a new user record to the system, the network administrator executes the command:

```
store(username, profile)
```

on one of the nodes (e.g. from the administration console or by logging on to one of the APs using administrator's local AP login). Herein, the profile is a list of authorizations. Principally, such a profile could be in an arbitrary suitable format. In this paper, we use the convenient attribute value pairs (AVPs) as they are typically used in the AAA systems today. The AVPs define the authentication method, the restrictions and session parameters (credentials, session duration, group membership, allowed access points, time constraints, etc). A typical profile hardly ever exceeds 10kB.

Similarly, the user delete is invoked as a DHT primitive:

```
delete(username)
```

This command removes the profile of the respective user from the overlay database.

The security issues concerning such requests are comparable to the respective access to an AAA database. These could be resolved e.g. by protecting the overlay traffic and by signing the requests with a unique administrator private key.

User Network Access

When a user accesses a COMPASS network, the user's mobile station (MS) and the solicited AP start the typical 802.1X authentication process. Within this process, at some point of time MS sends an `EAP Response/Identity` message containing the identity string. This is illustrated in Figure VI-19. AP_1 retrieves the corresponding user profile from the overlay by invoking an overlay lookup for this string as a key. On the receipt of the profile for the user, the AP_1 can continue the EAP conversation as defined in the profile acting as a 802.1X authenticator with a local 802.1X authentication server (AS). It verifies the access authorizations and applies the link authorizations from the profile to the established link. Depending on whether location privacy is supported (as e.g. in PEAP and EAP-TTLS), the first `EAP Response/Identity` message does not necessarily contain the "real" user login name. This is the same situation as for an AAA server which also needs to retrieve the right user profile. In this case the AP has to proceed with the requested EAP conversation until the point when it obtains the "real" user access identity (typically tunneled within the main EAP conversation). From here on it proceeds as described above.

Note that in every mutual authentication method, the identity used by the AP is an abstract identity of the whole overlay since from the user's point of view the network logically acts as one entity. Thus the overlay has to store the network credentials used in the network access phase. The exact nature of these depends on the used EAP authentication method and in the general case also has to be retrieved from the overlay first.

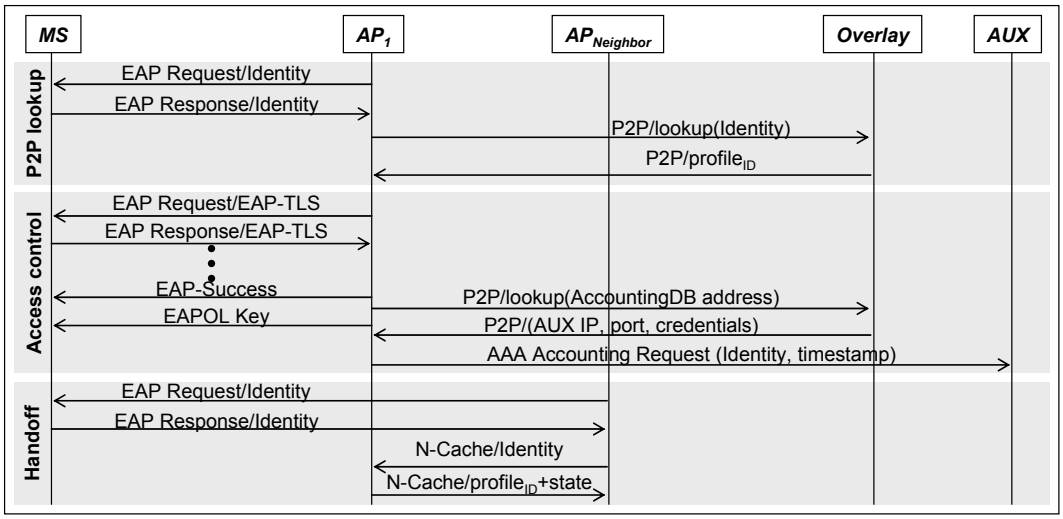


Figure VI-19 Message flow in the P2P overlay network and on the user link

Data Delivery

Principally the overlay can be used as a database or as a name resolution service. In both cases a host looking e.g. for a service `X` executes `lookup(X)` in the overlay. While in the first case the received answer is the value corresponding to the key `X` (as described for user profiles), in the second case the answer is the address of the service end point having the value of `X` (e.g. an URI). We demonstrate this with an example. Consider an access point searching for an update of its firmware. The key could be e.g. the string "newest firmware Linksys WRT54". In the first case, the overlay stores the complete firmware file. In the second case, the overlay stores the URL where to get the current firmware file; the AP could download the actual file in a subsequent step.

We call the first data delivery method an *integrated data delivery* while the second case is denoted as an *iterated data delivery*. Both approaches have their very particular pros and cons and it is difficult to uphold one particular approach in a general case.

The advantages of an integrated delivery are:

- Self-containing: simple implementation logic, no third party
- Overlay security applies to the delivered data – no relying on exterior security

The advantages of an iterated delivery are:

- Generally more efficient (direct transport layer transport)
- Smaller overlay database
- Permits storage on chosen, predestined hosts

We propose a mixed approach, trying to tie together the advantages of both delivery types. We define criteria by which an administrator (or an implementer) can decide which approach to use. Depending on the nature of the requested data, our overlay either directly holds the value or only stores the address where the latter can be retrieved from.

We namely use the data record size and data confidentiality need as criteria. Data with big records should use the iterated approach. Highly confidential and time-critical data should use the integrated delivery. Consequently, COMPASS directly stores user profile data (relatively small data records, highly confidential data) and its own management data (idem). General configuration data and the firmware update files are kept on exterior, dedicated file servers (public, non confidential data with very big records).

Auxiliary Nodes

The P2P architecture of our backend network naturally allows binding additional hosts (i.e. non APs) into the overlay. We provide support for such hosts which we call *auxiliary nodes*. These are explicitly optional but may be convenient in some cases.

These cases include databases storing big entries or highly dynamic non-volatile data. Typical examples of such databases would be session accounting data or firmware update files for the deployed equipment. The contents of such databases should not be stored within the overlay for efficiency and/or safety reasons.

The support for auxiliary nodes allows for easy extensions that might be necessary with the network growth. These are still optional because a similar result can be achieved using the iterated data delivery and explicit access to the retrieved address

Failure Management and Optimizations

We distinguish two major possible failures: the path failure (meaning that some of the nodes in the path are not available) and the end node failure (meaning that the zone database is not available). CAN provides different mechanisms to counter the impacts of such failures.

In our particular application, the path failure is not fatal. Given the stable wired networking we consider it a minor issue. By increasing the number of dimensions (`d`) of the CAN, the number of possible (and tried) paths can be increased thus reducing the probability of the overlay path failure.

If the overlay stores the only copy of the database, the failures of nodes holding a zone database are fatal. CAN provides two data replication methods. By using multiple realities or multiple different hash functions, the same zone database can exist on multiple nodes. In our scope, the use of such mechanisms is encouraged e.g. for the user DB.

We demonstrate this with an example of CAN with `k` realities. Let `f` of `n` nodes fail. Assuming that the overlay routing and delivery are still fully operational (i.e. all overlay paths still work), the probability that a stored (key, value) pair can be retrieved is:

$$p = 1 - \left(\frac{f}{n}\right)^k$$

Example: in a CAN with 4 realities, if ¼ of all nodes fail without path failures, there is still 99.6% probability that a user can still log in.

Discussion

CAN technology has been recently used in sensor networks [150]. Compared to sensors, the WLAN access points are powerful machines: usual 802.11 access points have about 16-32MB RAM and a MIPS CPU of about 150MHz and more. This principally allows the use of different P2P overlays. However the resource requirements related to our proposal need to be minimized since the access points are designed to be capable of serving the WLAN traffic, with some safety margins. Our proposal should not have impacts on the main access point behavior and thus needs to operate within this safety margin. Recent 802.11 APs with an embedded Linux OS show that this margin is sufficiently big for additional tasks.

As mentioned, the CAN itself has constant memory requirements. The management database mainly stores settings valid for every 802.11 AP. Its size is thus (almost) independent of the number of APs[1].

Given a typical profile size of some kilobytes, 30-100 user profiles per 802.11 AP can principally be stored without any impact on AP performance. The exact number mainly depends on the AP type, the average profile size and the parameters of the overlay. The user profile size can be further reduced by using group management. Overlay redundancy mechanisms can not be reasonably used when the overall number of APs is very low (say for up to 5-10 APs). When using redundancy mechanisms, the overall database size has to be multiplied by the redundancy factor `k`. For instance, in a network with 10 APs, 100 user profiles and redundancy factor `k=4`, the zone database on each AP is about `5kB*100*4/10=2MB`. These values seem to be realistic requirements for modern APs.

In Table VI-8, we compare different AAA-based access control architectures and COMPASS in terms of provided user mobility, administration complexity, network extensibility (complexity of extensions and scaling), and robustness (expressed by the worst case impact of a partial system failure).

Table VI-8 Comparative chart of user management methods in modern WLANs

Access Control Infrastructure	User mobility	Admin. complexity	Network extension		Partial system failure impact
			Complexity	Scaling	
Single central AS	Unlimited	Moderate	Easy	Bad	Fatal (SPF at AS)
Mult. ASs with a user DB per AS	Unlimited	Highest	Difficult	Good	Partial, no access for some user groups
Mult. ASs with a common user DB	Unlimited	High	Moderate	Bad	Fatal (SPF at DB)
APs with local user accounts	Limited	Low	Impossible	-	Partial, no access for local users
Central AP with user accounts	Unlimited	Moderate	Easy	Very bad	Fatal (SPF at AP)
COMPASS	Unlimited	Low	Automatic	Good	None or partial, adjustable over d and k

VI.6.6. COMPASS and User Roaming

As already explained above, an important aspect of the AAA usage today is the support for user roaming. In our access model we also presume that every user is known to at least one provider in the whole system. In this section, we describe COMPASS support for user roaming explicitly explaining how standard AAA roaming can be used with COMPASS.

Bidirectional AAA to COMPASS Roaming

In COMPASS every AP implements a very basic AAA server for compatibility reasons. Thus, AAA roaming can be used between COMPASS SPN and an SPN using standard AAA. AAA roaming is based on a realm that is extracted from the user name. NAI is typically used for that purpose. AAA roaming is unidirectional and needs different parameters at the server relaying the request (*proxying server*) from an AP and the end server. The relaying server typically uses a roaming configuration per realm storing a FQDN (fully qualified domain name), a destination port and a shared secret. For the end

[1] In any case, every AP is designed to be capable of storing its own configuration settings. Our approach results in the same behavior in the worst case. In the best case, one common configuration file is saved in the overlay and only differences are stored per AP.

server the configuration of a proxying server is equivalent to the configuration of any other client (FQDN, shared secret).

Since the AAA proxying is technically unidirectional, two proxying relationships have to be established for a mutual proxying. These are however technically different.

If a SPN using COMPASS with a common FQDN `compass.example.com` is to be interconnected with an SPN using AAA with FQDN `aaa.example2.com`, the configuration can be as presented in Table VI-9.

Table VI-9 Proxying configuration parameters in different interconnection cases

	COMPASS to AAA proxying		AAA to COMPASS proxying	
	COMPASS: destination server configuration	AAA server: new client configuration	AAA server destination server configuration	COMPASS: external AAA client configuration
Realm	example2.com	-	example.com	-
From/To FQDN	aaa.example2.com	compass.example.com	compass.example.com	aaa.example2.com
Shared secret	S_0	S_0	S_1	S_1
Configuration data stored in	Distributed COMPASS management DB	AAA server configuration file	AAA server configuration file	Distributed COMPASS management DB
Configuration data format	Key: Realm Value: (dest FQDN, port, shared secret)	-	-	Key: AAA client address Value: (shared secret)

The mentioned COMPASS FQDN can be either the FQDN of a special auxiliary host or a common round robin DNS name. While the latter can be used for AAA to COMPASS roaming, it can not be used for COMPASS to AAA roaming with the standard client authentication. The problem is that a client (i.e. AP) from the COMPASS SPN accessing the AAA server in the foreign SPN does not necessarily have the IP address to which the configured common COMPASS DNS name resolves. The round robin DNS resolves in a different IP address at each request and could thus resolve to any other address stored under the same DNS name. Therefore, in the case of COMPASS to AAA roaming, an auxiliary host has to be used unless an alternative AAA-client identity verification is used or some special treatment applied (one could e.g. resolve all IPs under the given DNS address and see if the requesting IP is in the list).

Apart from this detail, the usage of an auxiliary host vs. a common DNS name generally results in advantages and weaknesses from the robustness and security perspectives. Evidently, an auxiliary host represents a SPF for all roaming to/from the SPNs served over this auxiliary host. At the other hand, one auxiliary host might be considered as easier to secure by the local administration. Indeed, it is simpler to control and to limit access to one single host than to all nodes building the COMPASS overlay.

COMPASS to COMPASS Roaming

COMPASS natively and transparently supports the inter-AN roaming within an SPN. COMPASS nodes have access to user profiles within the distributed database and can locally verify the validity of the accessing users. The topological neighborhood and the caching accelerate this process.

The inter-SPN COMPASS to COMPASS roaming has to use some open inter-provider interface. According to our access model and since the participating nodes typically implement an AAA server, an AAA protocol is an obvious candidate. When AAA is used

as transport protocol for COMPASS to COMPASS SPN roaming, the interconnection can be achieved with or without auxiliary hosts (the round robin DNS problem can be easily circumvented within COMPASS), similarly to AAA-COMPASS roaming.

VI.6.7. Conclusion

In this chapter we discuss the decentralization of the control plane of an SPN. We think that such decentralization can be interesting for multiple potential 4G service provider types since it permits to conceive control plane architectures capable of accompanying the natural network development.

At the example of the currently typical AAA architectures we discuss the shortcomings of the centralized concepts in both big and small networks. From these shortcomings we derive the requirements on the control plane architecture and motivate our search for alternative solutions. We then present new paradigms and concepts that seem to be suitable candidates for a new control plane architecture. We namely retain the Distributed Hashtable (DHT) peer-to-peer technology and present the Content Addressable Network (CAN) as a particularly interesting candidate.

We concretize our ideas by integrating CAN's DHT technology with 802.1X access control. We develop an 802.11 system without any central element, thus supporting a natural scaling of the management infrastructure. As a truly distributed P2P architecture, our Configuration Management P2P Access Security System (COMPASS) inherits the scalability and fault tolerance of the existing DHT technologies and the compatibility of the 802.1X access control. It is thus able to fulfill the requirements defined above. COMPASS additionally features easy network extensibility, auxiliary host support and AAA-like user management. Moreover, by storing configuration and management settings in the overlay, COMPASS can also be used as network management infrastructure.

Finally, we show how COMPASS can be principally used in our global 4G system vision by supporting AAA-like user roaming from and to SPNs using COMPASS.

VI.7. Conclusion

Pursuing the main goal of providing different services over heterogeneous access networks within the same SPN, in this Chapter we have addressed the organization of the service provider networks.

We start with the adaptation issues in the network plane. We namely study the achievable flexibility of modern 802.11 ANs and present existing virtualization techniques to achieve such flexibility in a manner fulfilling our security requirements. We then study the adaptation capabilities of the core network's network plane and present two different virtualization approaches RESACO and MMQOS achieving a dynamic per-user network-to-user adaptation. In this scope we explicitly address the access security, access network choice, user-dependent service discovery and QoS-aware service access issues.

We then address the organization of the SPN's control plane motivating alternative approaches to the currently predominant centralized concepts. Our approach is an adaptable, scalable and self-organizing P2P control plane architecture. We introduce and discuss COMPASS, our proposal for the implementation of this idea in a 802.11 environment using 802.1X access control.

C H A P T E R V I I

Conclusion

VII.1. Summary of Contributions

Since no existing technology can resolve the access network problem as defined in Section III.1, a system allowing for a flexible choice of a transmission technology is a seductive concept. Yet the handling of such a heterogeneous system is a very complex issue. The real problem is not the raw data transport per se but the organization of *how* the data are transported over different infrastructures owned by different authorities. The dilemma is that homogenizing too much is unrealistic while preserving complete heterogeneity lacks integration. A reasonable tradeoff is necessary between the standardization (i.e. homogenization) on one hand and flexibility (i.e. diversity) on the other.

To achieve such an optimal tradeoff, the system focus must be set on the user, as opposed to service or provider. The problem with the classical service-oriented design is the limitation of the designed system to the anticipated services only. The problem with the provider-oriented design is the designed closed network architecture and the high service cost for the user. User-oriented design primarily implies that every service can be used by every user independently of the underlying technology, of the provider owning the infrastructure and of whom the user has signed the contract with. In this view, the service contract can actually be reduced to one single practical purpose: preliminary trust

establishment. If this constraint can be bypassed by other means, no traditional service contracts are necessary. This is e.g. true for a credit card issuer acting as a virtual operator.

In this work we applied user-oriented design principles to the different problem areas of the next generation mobile communications system. We apply the user-oriented design in the mentioned technology-opportunistic approach to 4G. By that means, we reduce the system complexity to a fairly simple access model concentrating on a user accessing a service over a serving provider with the help of the home provider. Herein, the latter is used as a commonly trusted entity. The service provider acts as a service access network. The location of the service itself is undetermined. This view permits us to rethink the complexity in terms of interfaces. We namely identify the user-provider interface and the provider-provider interface.

We then apply the virtualization and the cross-layer design to overcome the access network heterogeneity and the problems resulting from the layered approach respectively.

VII.1.1.4G Network Access Contributions

On the user-provider interface, we define the logical access problem. We then use the virtualization as the mechanism achieving a reasonable tradeoff between standardization and flexibility. With virtualization, only the interface but not the actual implementation is specified. Furthermore, different implementations are explicitly supported and can be used interchangeably if the interface specification is respected.

We use this approach on the user-provider interface and propose an alternative access model using a standardized logical access means along with a technology-specific data transport. This dual approach is particularly interesting since it provides more flexibility and can be more efficient by applying the cross-layer design and avoiding the “layered” delays. It also explicitly covers another important aspect of the user-oriented design: the usage comfort. The common (technology-independent) logical signaling part permits a technology-independent network and service discovery and can thus establish a decision base for the choice of the network to use. Through our early discovery, we provide a zero pre-configuration access to each conform roaming network and thus implement the aforementioned requirement for system-wide user service access. Because of practical constraints, and notably the current orientation of most attacks on IP, this is extremely difficult to achieve in a secure way in pure All-IP architectures.

We proposed a working implementation of our ideas in the context of modern 802.11 WLANs using 802.1X access control. We generalized the EAP usage in this scope and designed EAP-SIG, a protocol capable of such technology-independent signaling transport. Since we are conformly using the existing standards, our approach can be used with the existing EAP equipment. Because of the medium-independence of EAP, this approach is principally applicable in the 4G scope.

We also addressed different mobility-related issues with the WLANs in the 4G context. We namely studied user roaming support in 802.11 WLANs and defined requirements that later became common best practices. We also studied the IP micromobility integration in the WLANs. Although the latter seems straightforward following the layered approach, we identified different potential problems and proposed an alternative integration scheme that – without changes to the used mechanisms – can resolve some of the identified issues.

VII.1.2. Contributions to SPN Architectures

In spite of the user-oriented design, we need to attribute the necessary attention to the provider's need for management of the deployed architectures. Without a reliable control, no reasonable business scenarios are possible. However, the management of the heterogeneous architectures represents one of the key 4G problems.

To reflect the heterogeneity and to identify a minimal common part, we separated the provider infrastructure in access networks implementing the transmission technologies and the common IP-based SPN core containing user services. We then study the requirements on the access networks in terms of management and configurability and the needs for the flexibility within such an infrastructure.

Analogously, the virtualization provides the necessary adaptability in this context. We study this in a practical, fully operational 802.11 campus network environment. We designed and installed a network architecture that allows the instantiation of different environments (including the L2 parameter adaptation) upon one and only deployed physical infrastructure. The automation in this scope can provide "one click virtual network creation", if necessary.

We then addressed the SPN adaptation issues proposing two alternative schemes to provider-internal organization. Both our approaches aim to maximize the service provision flexibility for the providers remaining transparent for the users.

We designed a highly adaptable SPN solely parameterized by the user profile of the accessing roaming user. In our approach, both the visible service environment and the actually accessible services change depending on the accessing user even if all SPN access occurs over the same access network. We implemented this approach at an example of a 802.11 network with user roaming between two universities. Our prototype implementation runs on an old portable laptop PC and thus seems suitable for the embedded access point implementation.

We designed an SPN architecture pushing control elements beyond the network edge, i.e. in the user terminal. This is possible due to the usage of the tamper-resistant modules like smartcards. The smartcard provides a virtual interface. Using this interface, we hide the implementation complexity from the user terminal providing once again a common open interface. This approach guarantees a stable service environment through the proposed SoC concept and the internal dynamic connections from the smartcard to services in different provider networks. The smartcard implementation permits to optimally choose the next available service and thus avoids complicated routing as in the case of current VPN modules. This approach can be seen as a generalization of the (U)SIM module known from GSM and UMTS.

Finally, we studied the organization of the control plane of a SPN. We identified insuffficiences with the current centralized approaches. The main problems are the bad scalability and the presence of a single service of failure. We show that such approaches can not be reasonably used in both very large and very small SPNs. Leveraging the recent advances in the P2P networking, we proposed an original alternative control plane organization. We introduced COMPASS, our configuration and management P2P-based system providing a secure service access. COMPASS scales automatically with the network growth. COMPASS also avoids any SPF. We presented the main mechanisms used in COMPASS and studied its robustness. We also studied interoperation scenarios between COMPASS and other proposed architectures.

VII.2. Directions for Future Research

In a complicated heterogeneous system consisting of several technologies and different authorities, there will always be open issues and need for further improvements. Except, since we do not study the fourth remaining interface (user-user interface), our contributions do not yet address a whole area of open issues like e.g. end-to-end QoS.

Concentrating on the areas which we addressed, the most of our contributions clearly need further study, especially field experiments within an operational environment.

We conclude this document by identifying the following interesting points within our started work which seem to be valuable topics for further research and development:

- The prototype implementation of EAP-SIG concept can be further developed. This will help to understand to which degree the inherited characteristics of EAP represent a practical limitation. Principally, EAP-SIG can be extended to provide a virtual interface which can then be used by the existing frameworks like e.g. UPnP, etc. The main advantage of such approach would be a smooth integration. However, we believe that the signaling channel should be intentionally kept limited since it naturally renders attacks over this channel less attractive.
- Alternatively, a new protocol should be taken into consideration to implement a generic signaling transporter. Such a protocol would not provide an immediate compatibility like EAP does. However, it could be designed based on a different communications model thus allowing for easier signaling start and not suffering from client-server limitations. The efforts in the recently founded IEEE 802.21 Working Group seem interesting in this scope.
- Our generic two channel model should be studied for suitability for a looser AN and SPN discovery. For instance, the signaling channel established over one AN could be used to discover other ANs and/or parameters of other ANs. This seems to be a viable approach for UMTS/WLAN interworking.
- Since our RESACO implementation, real 802.11 access points with an embedded Linux OS have appeared on the market. Thus, this approach should be implemented on a real access point and tested in an operational environment.
- Analogously, the progress in the smartcard design and the development of new Java-based IP smartcards opens a principle possibility for a prototype implementation. This would the practical verification of the analytically estimated performance values.
- Although both proposed provider architectures RESACO and MMQOS base on the integration of standard protocols and mechanisms which are thus expected to perform adequately, the designed architectures might need a robustness and scalability study supporting user mobility scenarios. This could be carried out by the means of a simulation or a test bed implementation.
- COMPASS is an interesting concept which deserves a further study. Although it seems viable, we need performance values. Since some implementations of DHTs exist by now, a real platform can be developed e.g. in a 802.11 test environment. An alternative to DHT can be found and studied for COMPASS.
- There seems to be some so far unused potential in the study of the synergetic effects between our contributions. This namely includes the usage of EAP-SIG within RESACO, MMQOS and COMPASS. Such usage is principally straightforward by design. However, the recent development of EAP smartcards opens some new possibilities. Eventually, a complete architecture implying the most interesting approaches of EAP-SIG, MMQOS or RESACO with COMPASS in the distributed control plane could be a new interesting proposal.

References

[1] K. Annan, Secretary-General of the UN, "Message on the Opening of World Space Week", http://www.oosa.unvienna.org/wsw/2000/sgtext_E.html, New York, October 4, 2000.

[2] C. Shirky, "Half the World", http://www.shirky.com/writings/half_the_world.html, September 3, 2002.

[3] International Telecommunications Union, Telecom Development Sector (ITU-D), Information and Communication Technology (ICT), Free Statistics, "Main telephone lines, subscribers per 100 people: 2001", http://www.itu.int/ITU-D/ict/statistics/at_glance/main01.pdf.

[4] International Telecommunications Union, Telecom Development Sector (ITU-D), Information and Communication Technology (ICT), Free Statistics, "Main telephone lines, subscribers per 100 people: 2001", http://www.itu.int/ITU-D/ict/statistics/at_glance/cellular01.pdf.

[5] GSM Association – GSM World, GSM Technology, http://www.gsmworld.com/technology/gsm.shtml.

[6] Central Intelligence Agency, "World's Fact Book", http://www.cia.gov/cia/publications/factbook/.

[7] International Telecommunications Union, Telecom Development Sector (ITU-D), Information and Communication Technology (ICT), Free Statistics, "Internet

indicators: Hosts, Users and Number of PCs: 2003“, http://www.itu.int/ITU-D/ict/statistics/at_glance/Internet03.pdf.

[8] ICFA-SCIC Monitoring Working Group, “ICFA-SCIC Network Monitoring Report”, International Committee for Future Accelerators (ICFA) – Standing Committee on Inter-Regional Connectivity (SCIC), January, 2004, http://pcstats.cern.ch/icfa-scic/Documents/20040206-NetwMonit-A4.pdf.

[9] F. Postogna, C. Fonda, E. Canessa, G. O. Ajayi and S. Radicella, “Wireless Networking in Africa”, Linux Journal #56 Dec 1998 issue.

[10] A. Pentland, R. Fletcher, A. Hasson, “DakNet: Rethinking Connectivity in Developing Nations”, IEEE Computer, vol. 37, issue 1, pp. 78-83, January 2004.

[11] The Official Bluetooth Wireless Info Site, http://www.bluetooth.com.

[12] IEEE Standard 802.15.1, “Wireless Medium Access Control (MAC) and Physical Layer (PHY) Specifications for Wireless Personal Area Networks (WPANs)”, June 2002.

[13] ZigBee Alliance, http://www.zigbee.org.

[14] IEEE Standard 802.11, “Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications”, 1999 Editions, 1999.

[15] ETSI, “HIPERLAN/2 Standard”, http://portal.etsi.org/bran/kta/Hiperlan/hiperlan2.asp.

[16] IEEE Standard 802.16, “Air Interface for Fixed Broadband Wireless Access Systems”, December 2001.

[17] ETSI Broadband Radio Access Networks, http://portal.etsi.org/bran/Summary.asp.

[18] IEEE 802.16 Working Group, http://www.ieee802.org/16.

[19] GSM Association, 3GSM Platform, http://www.gsmworld.com/technology/3g/index.shtml.

[20] CDMA Development Group, Technology: 3G - cdma2000, http://www.cdg.org/technology/3g.asp.

[21] IEEE 802.20 Working Group, “Mobile Broadband Wireless Access”, http://www.ieee802.org/20/.

[22] A. S. Tanenbaum, “Computer Networks”, ISBN 0130661023, Prentice Hall PTR, 4th edition, August, 2002.

[23] D. Raychaudhuri, “4G Network Architectures: WLAN Hot-Spots, Infostations and beyond...”, IEEE PIMRC 2002 Keynote Talk, Lisbon, Portugal, September, 2002.

[24] G. E. Moore, “Cramming more components onto integrated circuits“, Electronics, vol. 38, number 8, April, 1965.

[25] J. D. Day and H. Zimmerman, “The OSI reference model”, In Proceedings of the IEEE, vol. 71, pages 1334--1340, 1983.

[26] Q. Rahman and M. Ibnkahla, Chapter 6, “Mobile Communications: Technologies and Challenges”, in M. Ibnkahla Ed., “Signal Processing for Mobile Communications Handbook”, ISBN 084931657X, CRC Press, 2004.

[27] International Telecommunications Union, ITU Radio-Frequency Spectrum Publications, http://www.itu.int/publications/main_publ/frequency.html.

[28] E. van Damme, "The European UMTS-auctions", European Economic Review, pp. 846-858, vol. 46, issue 4-5, May 2002.

[29] J. Geier, "Duelling with Microwave Ovens", http://www.wi-fiplanet.com/tutorials/article.php/3116531.

[30] RNRT Project INFRADIO, http://rp.lip6.fr/infradio/.

[31] C. Perkins, Ed. "IP Mobility Support for IPv4," RFC 3344, IETF, August 2002.

[32] J. Rosenberg et al., "SIP: Session Initiation Protocol", RFC 3261, IETF, June 2002.

[33] D. Tiang, M. Baker, "Analysis of a Local-Area Wireless Network", ACM MOBICOM, Boston, USA, 2000.

[34] A. Campbell et al., "Comparison of IP Micromobility Protocols," IEEE Wireless Communications, pp. 72-82, February 2002.

[35] S. Das et al., "IDMP: An Intradomain Mobility Management Protocol for Next-Generation Wireless Networks," IEEE Wireless Communications, pp. 38-45, June 2002.

[36] R. Ramjee et al., "HAWAII: A Domain-Based Approach for Supporting Mobility in Wide-area Wireless Networks," Proc. IEEE Int'l. Conf. Network Protocols, 1999.

[37] P. De Silva and H. Sirisena, "A Mobility Management Protocol for IP-Based Cellular Networks," IEEE Wireless Communications, pp. 31-37, June 2002.

[38] B. Schneier, "Applied Cryptography", ISBN 0471117099, Wiley, 2nd edition, October 1995.

[39] A. Kerckhoffs, "La cryptographie militaire", Journal des Sciences Militaires, vol. 9, p. 12, January 1883.

[40] Y. Bai, H. Kobayashi, "Intrusion Detection System: Technology and Development", in Proc. Of 17th AINA, p. 710, 2003.

[41] D. J. Gooch, S. D. Hubbard, M. W. Moore and J. Hill, "Firewalls – evolve or die", BT Technol J, pp. 89-98, vol. 19, No. 3, July 2001.

[42] Y. Xu, H. C. J. Lee, "A Source Address Filtering Firewall to Defend Against Denial of Service Attacks", in Proc. 60th Vehicular Technology Conference, Los Angeles, CA, USA, September 2004.

[43] D. Patiyoot, S. J. Shepherd, "Cryptographic Security Techniques for Wireless Networks", ACM SIGOPS Operating System Review, pp. 36-50, vol. 33, issue 2, April 1999.

[44] S. Ravi, A. Raghunathan and N. Potlapally, "Securing Wireless Data: System Architecture Challenges", ACM ISSS'02, Kyoto, Japan, October 2002.

[45] Wireless Network Visualization Project, ITTC, University of Kansas, http://www.ittc.ku.edu/wlan/.

[46] Globalstar Corporation, http://www.globalstar.com/.

[47] Iridium Satellite Solutions, http://www.iridium.com/.

[48] H. Poincaré, "The Foundations of Science", p. 129, ISBN 0819123188, University Press of America, 1982.

[49] J. M. Pereira, "Fourth Generation: Now, it is Personal!", in Proc. IEEE PIMRC, pp. 1009-1016, vol. 2, London, UK, September 2000.

[50] ITU-R World Radiocommunication Conference, http://www.itu.int/ITU-R/conferences/wrc/index.asp.

[51] B.G. Evans and K. Baughan, "4G Visions", IEEE Electronics & Communications Engineering Journal, pp. 293-303, December 2000.

[52] Y. Raivio, "4G – Hype or Reality", IEE 3G Mobile Communication Technologies, pp. 346-350, Conference Publication No. 477, March 2001.

[53] U. Varshney and R. Jain, "Issues in Emerging 4G Wireless Networks", IEEE Computer, pp. 94-96, June 2001.

[54] A. Bria, F. Gessler, O. Queseth, R. Stridh, M. Unbehaun, J. Wu, J. Zander and M. Flament, "4th-Generation Wireless Infrastructures: Scenarios and Research Challenges", IEEE Personal Communications, pp. 25-31, December 2001.

[55] V. Gupta and S. Gupta, "KSSL: Experiments in Wireless Internet Security", in Proc. Wireless Communications and Networking Conference, pp. 860-864, March 2002.

[56] L. Becchetti, F. D. Priscoli, T. Inzerillli, P. Mähönen, L. Muñoz, "Enhancing IP Service Provision over Heterogeneous Wireless Networks: A Path towards 4G", IEEE Communications Magazine, pp. 74-81, August 2001.

[57] W. Emmerich, "Engineering Distributed Objects", ISBN 0471986577, John Wiley & Sons; 1st edition, June, 2000.

[58] J. Al-Muhtadi, D. Mickunas and R. Campbell, "A Lightweight Reconfigurable Security Mechanism for 3G/4G Mobile Devices", IEEE Wireless Communications, pp. 60-65, April 2002.

[59] T. Otsu, I. Okajima, N. Umeda and Y. Yamao, "Network Architecture for Mobile Communications Systems Beyond IMT-2000", IEEE Personal Communications Magazine, pp. 31-37, October 2001.

[60] H. Yumiba, K. Imai and M. Yabusaki, "IP-Based IMT Network Platform", IEEE Personal Communications Magazine, pp. 18-23, October 2001.

[61] W. Kellerer, H.-J. Vögel, K.-E. Steinberg, "A Communication Gateway for Infrastructure-Independent 4G Wireless Access", IEEE Communications Magazine, pp. 126-131, March 2002.

[62] J. Zhang, J. Li, S. Weinstein, N. Tu, "Virtual Operator based AAA in Wireless LAN Hot Spots with Ad-hoc Networking Support", ACM Mobile Computing and Communications Review, pp. 10-21, vol. 6, No. 3, July 2002.

[63] D. Forsber, Y. Ohba, B Pati, H. Tschofenig, A. Yegin, "Protocol for carrying Authentication for Network Access", IETF PANA Working Group Draft, work in progress, March 2003.

[64] J. Loughney, Ed., M. Nakhjiri, C. Perkins, R. Koodli, "Context Transfer Protocol", draft-ietf-seamoby-ctp-11.txt, approved IETF draft, work in progress, August 2004.

[65] P. Bahl, A. Balachandran and S. Venkatachary, "Secure Wireless Internet Access in Public Places", IEEE ICC 2001, Finland, June 2001.

[66] A. Friday, M. Wu, S. Schmid, J. Finney, K. Cheverst and N. Davies, "A Wireless Public Access Infrastructure for Supporting Mobile Context-Aware IPv6 Applications", in Proc. ACM 1st Workshop on Wireless Mobile Internet, pp. 11-18, Rome, Italy, July 2001.

[67] L. Dell'Uomo, E. Scarrone, "The Mobility Management and Authentication / Authorization mechanisms in Mobile Networks beyond 3G", IEEE Personal, Indoor and Mobile Radio Communications, pp. C44-C48, vol. 1, September 2001.

[68] S.-L. Tsao and C.-C. Lin, "Design and Evaluation of UMTS-WLAN Interworking Strategies", in Proc. IEEE 56th VTC, Vancouver, Canada, September 2002.

[69] T. Zhang and P. Agrawal, J.-C. Chen, "IP-Based Base Stations and soft Handoff in All-IP Wireless Networks", IEEE Personal Communications Magazine, pp. 24-30, October 2001.

[70] A. Misra, S. Das, A. Dutta and A. McAuley, S. K. Das, "IDMP-Based Fast Handoffs and Paging in IP-Based 4G Mobile Networks", IEEE Communications Magazine, pp. 138-145, March 2002.

[71] P. Ginzboorg, "Seven Comments on Charging and Billing", Communications of the ACM, pp. 89-92, vol. 43, No. 11, November 2000.

[72] M. Peirce, "Multi-Party Electronic Payments for Mobile Communications", PhD Thesis, Department of Computer Science, University of Dublin, Trinity College, October 2000.

[73] E. Rosen, A. Viswanathan, R. Callon, "Multiprotocol Label Switching Architecture", RFC 3031, IETF, January 2001.

[74] R. Braden, D. Clark, S. Shenker, "Integrated Services in the Internet Architecture: an Overview", RFC 1633, IETF, June 1994.

[75] S. Blake, D. Black, M. Carlson, E. Davies, Z. Wang, W. Weiss, "An Architecture for Differentiated Services", RFC 2475, IETF, June 1998.

[76] S. Kent, R. Atkinson, "Security Architecture for the Internet Protocol", RFC 2401, IETF, November 1998.

[77] T. Dierks, C. Allen, "The TLS protocol version 1.0", RFC 2246, IETF June 1999.

[78] GSM 11.11, "Digital Cellular Telecommunication System (Phase 2+), Specification of the Subscriber Identity Module – Mobile Equipment (SIM-ME) Interface".

[79] 3GTS 33.102 Release 99, "3GPP: Technical Specification Group (TSG), 3G Security: Security Architecture".

[80] IEEE Standard 802.1X, "Port-Based Network Access Control", IEEE, June 2001.

[81] IEEE Draft 802.11i, "Draft Supplement to IEEE Std 802.11. Part 11: Specifications for Enhanced Security", work in progress.

[82] IEEE Standard 802.11F, "Trial-Use Recommended Practice for Multi-Vendor Access Point Interoperability via an Inter-Access Point Protocol Across Distribution Systems Supporting IEEE 802.11 Operation," July 2003.

[83] IEEE Draft 802.11e, "Draft Supplement to Standard For Telecommunications and Information Exchange Between Systems – LAN/MAN Specific Requirements – Part 11: Wireless Medium Access Control (MAC) and Physical Layer (PHY) specifications: Medium Access Control (MAC) Enhancements for Quality of Service (QoS)", work in progress, Februray 2003.

[84] J.D. Case, M. Fedor, M.L. Schoffstall, J. Davin, "Simple Network Management Protocol (SNMP)", RFC 1157, IETF, May 1990.

[85] D. Durham, Ed., J. Boyle, R. Cohen, S. Herzog, R. Rajan, A. Sastry, "The COPS (Common Open Policy Service) Protocol", RFC2748, Internet Society , January 2000.

[86] R. Yahalom, B. Klein, Th. Beth, "Trust Relationships in Secure Systems – A Distributed Authentication Perspective", in Proc. of IEEE ComSoc Symposium on Research in Security and Privacy, pp. 150-164, Oakland, CA, USA, May 1993.

[87] H. Schulzrinne, E. Wedlund, "Application-Layer Mobility Using SIP", ACM Mobile Computing and Communications Review, vol. 4, No. 3, pp. 47-57.

[88] UPnP Forum, http://www.upnp.org.

[89] OSGi Alliance, http://www.osgi.org.

[90] IETF Authentication, Authorization, Accounting (AAA) Working Group, http://www.ietf.org/html.charters/aaa-charter.html.

[91] C. Rigney, S. Willens, A. Rubens, W. Simpson, "Remote Authentication Dial-In User Service (RADIUS)", RFC 2865, IETF, June 2000.

[92] J. Hill, "An Analysis of the RADIUS Authentication Protocol", November 2001, http://www.untruth.org/~josh/security/radius/radius-auth.html.

[93] P. Calhoun, J. Loughney, E. Guttman, G. Zon, J. Arkko, "Diameter Base Protocol", RFC 3588, IETF, September 2003.

[94] B. Aboba, L. Blunk, J. Vollbrecht, J. Carlson, H. Levkowetz, Ed., "Extensible Authentication Protocol (EAP)", RFC 3748, IETF, June 2004.

[95] B. Aboba, D. Simon, "PPP EAP/TLS Authentication Protocol", RFC 2716, IETF, October 1999.

[96] IETF Extensible Authentication Protocol (EAP) Working Group, http://www.ietf.org/html.charters/eap-charter.html.

[97] D. Stanley, J. Walker, B. Aboba, "EAP Method Requirements for Wireless LANs", IETF draft, work in progress, draft-walker-ieee802-req-04.txt, August 2004.

[98] C. Rigney, W. Willats, P. Calhoun, "RADIUS Extensions", RFC 2869, IETF, June 2000.

[99] GET AUTHENTIS Project, http://www.enst.fr/~hecker/authentis.

[100] K. P. Bosworth and N. Tedeschi, "Public Key Infrastructures – the Next Generation", BT Technology Journal, vol. 19, No. 3, July 2001.

[101] Meeting of the Scientific Council of GET, ENST, Paris, France, February 2003.

[102] W. Simpson, Ed., "The Point-to-Point Protocol (PPP)", IETF RFC 1661, July 1994.

[103] K. Ahmavaara, H. Haverinen and R. Pichna, "Interworking architecture between 3GPP and WLAN systems", IEEE Communications, pp. 74-81, vol. 41, no. 11, November 2003.

[104] H. Dobbertin, "The Status of MD5 After a Recent Attack", RSA Laboratories' CryptoBytes, Volume 2, Number 2, 1999, ftp://ftp.rsasecurity.com/pub/cryptobytes/crypto2n2.pdf.

[105] X. Wang, X. Lai, D. Feng, H. Yu, "Collisions for Hash Functions MD4, MD5, HAVAL-128 and RIPEMD", CRYPTO 2004, Santa Barbara, CA, USA, August, 2004.

[106] B. Aboba, M. Beadles, "The Network Access Identifier", RFC 2486, IETF, January 1999.

[107] R. Droms, "Dynamic Host Configuration Protocol", RFC 2131, IETF, March 1997.

[108] BugTraq Mailing List, Symantec Corporation, http://www.securityfocus.com/.

[109] W. Simpson, "PPP Challenge Handshake Authentication Protocol (CHAP)", RFC 1994, IETF, August 1996.

[110] A. Misra, S. Das, A. McAuley, S. K. Das, "Autoconfiguration, registration, and mobility management for pervasive computing", pp. 24-31, IEEE Personal Communications, August 2001.

[111] A. Hecker, H. Labiod, "An Efficient Micromobility Implementation for 802.1X WLANs", IEEE 15th PIMRC, September 2004, Barcelona, Spain.

[112] J. Kempf, "Dormant host mode alerting ('IP Paging') problem statement," RFC 3132, IETF, June 2001.

[113] G. Judd and P. Steenkiste, "Fixing 802.11 access point selection", in Proc. ACM SIGCOMM 2002 Poster session, August 2002.

[114] F. Adrangi, V. Lortz, F. Bari, P. Eronen, M. Watson, "Mediating Network Discovery in the Extensible Authentication Protocol (EAP)", Internet draft, work in progress, IETF, draft-adrangi-eap-network-discovery-00.txt.

[115] I. Guardini, E. Demaria, J. Bournelle, M. Laurent-Maknavicius, "MIPv6 Authorization and Configuration based on EAP", Internet draft, work in progress, IETF draft-giaretta-mip6-authorization-eap-01.txt, July 2004.

[116] S. Pack and Y. Choi, "Fast Inter-AP Handoff using Predicitve-Authentication Scheme in a Public Wireless LAN," IEEE Networks 2002, Atlanta, USA, August 2002.

[117] A. Mishra, M. H. Shin, N. L. Petroni, T. C. Clancy, W. Arbaugh, "Proactive Key Distribution Using Neighbor Graphs", IEEE Wireless Communications, pp. 26-36, February 2004.

[118] S. Glass, T. Hiller, S. Jacobs, C. Perkins, "Mobile IP Authentication, Authorization, and Accounting Requirements," RFC 2977, IETF, October 2000.

[119] T. Noël, N. Montavont, P. Bertin, "Mobilité IPv6 et WLAN: Expérimentation et évaluation à l'échelle d'un campus", actes de DNAC 2002, http://www-rp.lip6.fr/dnac/7.1-noel-article.pdf, Paris, December 2002.

[120] J. Korhonen, "Performance Implications of the Multi Layer Mobility in a Wireless Operator Networks", 4th Berkeley-Helsinki Ph.D. Student Workshop on Telecommunication Software Architectures, Berkeley, CA, USA, June, 2004.

[121] C. Chiollaz, G. LeGrand, A. Hecker, F. Springinsfeld, S. Naqvi, Y. Denef, "Politique de sécurité pour une communauté ouverte dans un espace ouvert délimité", Délivrable D3a RNRT INFRADIO, INFRES, ENST, Paris, France, 2004.

[122] G. LeGrand, A. Hecker, F. Springinsfeld, "Architecture flexible de réseau sans fil WiFi sécurisé", SAR '04, La Londe, France, June 2004.

[123] IEEE Standard 802.1Q, "Virtual Bridged Local Area Networks", May 2003.

[124] Wi-Fi Alliance, "Wireless ISP Roaming (WISPr) – Best Current Practices", February 2003.

[125] K. Hamzeh, G. Pall, W. Verthein, J. Taarud, W. Little, G. Zorn, "Point-to-Point Tunneling Protocol", IETF RFC 2637, July 1999.

[126] M. Riguidel, "The Scientific Challenges of SEINIT Integrated Project – Security of Ambient Intelligence", White Paper, IST SEINIT Deliverable D1.0, WP1, February 2004.

[127] K. Eustice, L. Kleinrock, S. Markstrum, G. Popek, V. Ramakrishna, P. Reiher, "Securing Nomads: the Case for Quarantine, Examination, and Decontamination", in Proc. of ACM Workshop on New Security Paradigms, Ascona, Switzerland, 2003.

[128] GET RESACO Project, http://www.enst.fr/~hecker/resaco.

[129] The Click Modular Router Project, http://www.pdos.lcs.mit.edu/click/.

[130] Intel Software for UPnP Technology, http://www.intel.com/technology/UPnP/.

[131] D. C. Verma, "Policy-Based Networking Architecture and Algorithms", New Riders Publishing, Indianapolis, Indiana, USA, November 2000.

[132] W. Laouiti, K. Cordoso, A. Hecker, M. Genet, B. Jouaber, H. Labiod, D. Zeghlache, "RESACO: An Open and Programmable Multi-Domain Platform for Cooperative and Auto-Configurable Networks", ICWN 04, June 2004, Las Vegas, Nevada, USA.

[133] Meeting of the Scientific Council of GET, ENST Bretagne, Brest, France, January 2004.

[134] RNRT MMQOS project, http://www.enst.fr/~hecker/mmqos/.

[135] P. Urien, A. Tizraoui, M. Loutrel, K. Lu, "Integration EAP in SIM-IP smartcards", Workshop ASWN, 2002.

[136] M. Montgomery, A. M. Ali and H. K. Lu, "Secure Network Card – Implementation of a Standard Network Stack in a Smart Card", Six Smart Card Research and Advanced Application IFIP Conference, France, August 2004.

[137] H. Labiod, R. Duffau, "KMS: a key management system for multi-provider interconnected Wi-Fi WLANs", IEEE GLOBECOM '04, Dallas, TX, USA, December 2004.

[138] SUN Java Card 2.2.1 Platform Specification, http://java.sun.com/products/javacard/specs.html.

[139] TUN/TAP Project, Virtual Point-to-Point(TUN) and Ethernet(TAP) devices, http://vtun.sourceforge.net/tun/.

[140] P. Urien and M. Badra, M. Danjinou, "EAP-TLS Smartcards, from Dream to Reality", 4th Workshop on Applications and Services in Wireless Networks, Boston University, Boston, Massachusetts, USA, August 8-11, 2004.

[141] J. Tourrilhes, "L7 mobility: A Framework for Handling Mobility at the Application Level", in Proc. IEEE PIMRC, Barcelona, Spain, September, 2004.

[142] A. Goldberg, R. Buff, A. Schmitt, "A Comparison of HTTP and HTTPS Performance", Computer Measurement Group, CMG98, December 1998.

[143] V. Gupta, S. Gupta, S. Chang, "Performance Analysis of Elliptic Curve Cryptography for SSL", in Proc. ACM WiSe '02, Atlanta, Georgia, USA, September 2002.

[144] A. Levi, E. Savas, "Performance Evaluation of Public-Key Cryptosystem Operations in WTLS Protocol", in Proc. IEEE ISCC, Antalya, Turkey, July 2003.

[145] L. Bernard, "Integration of SIP and COPS protocols", Master thesis, University of Paris VI, 2002.

[146] K. Cardoso, M. G. Genet, D. Zeglache, "COPS Based Management for the UMTS multimedia domain", IEEE WPCM, October 2002.

[147] S. Ratnasamy, P. Francis, M. Handley, R. Karp, S. Shenker, "A Scalable Content-Addressable Network , Proceedings of the 2001 Conference on Applications, Technologies, Architectures, and Protocols for Computer Communications.

[148] I. Stoica, R. Morris, D. Liben-Nowell, D. Karger, M. F. Kaashoek, F. Dabek, H. Balakrishnan, "Chord: A Scalable Peer-to-peer Lookup Service for Internet Applications", Proceedings of the 2001 Conference on Applications, Technologies, Architectures, and Protocols for Computer Communications.

[149] A. Rowstron and P. Druschel, "Pastry: Scalable, Decentralized Object Location and Routing for Large-Scale Peer-to-Peer Systems", in Proc. IFIP/ACM International Conference on Distributed Systems Platforms (Middleware), 2001.

[150] H.-J. Hof, E.-O. Blaß, T. Fuhrmann and M. Zitterbart, "Design of a Secure Distributed Service Directory for Wireless Sensornetworks", EWSN 2004, Berlin.

[151] Linksys WRT54GS Wireless Access Point, ftp://ftp.linksys.com/datasheet/wrt54gs_ds.pdf.

[152] J. R. Walker, "Unsafe at any key size; An analysis of the WEP encapsulation", IEEE document 802.11-00/362, October 2000.

[153] N. Borisov, I. Goldberg, D. Wagner, "Intercepting Mobile Communications: The Insecurity of 802.11", in Proc. ACM SIGMOBILE, pp. 180-188, Rome, Italy, July 2001.

[154] W. A. Arbaugh, N. Shankar, Y. C. J. Wan, "Your 802.11 Wireless Network has No Clothes", in Proc. 1st IEEE International Conference on Wireless LANs and Home Networks, http://www.cs.umd.edu/~waa/wireless.pdf, March 2001.

[155] S. Fluhrer, I. Mantin, A. Shamir, "Weaknesses in the Key Scheduling Algorithm of RC4", in Proc. 8th Annual Workshop on Selected Areas in Cryptography, August 2001.

[156] A. Stubblefield, J. Ioannidis and A. Rubin, "Using the Fluhrer, Mantin and Shamir Attack to Break WEP", AT&T Labs Technical Report TD-4ZCPZZ, August 2001.

[157] A. Mishra, W. A. Arbaugh, "An Initial Security Analysis of the IEEE 802.1X Standard", Technical Report CS-TR-4328, Department of Computer Science, University of Maryland, http://www.cs.umd.edu/~waa/1x.pdf, February 2002.

Related Publications

Journal Papers

[1] A. Hecker, H. Labiod, G. Pujolle, H. Afifi, A. Serhrouchni and P. Urien, "*A New Access Control Solution for a Multi-provider Wireless Environment*", to appear in the Telecommunication Systems Journal, Springer (Kluwer Academic Publishers), 2005.

International Refereed Conference

[2] A. Hecker, H. Labiod, A. Serhrouchni, "*Authentis: Through Incremental Authentication Models to Secure Interconnected Wi-Fi WLANs*", IEEE ASWN 2002, Paris, France.

[3] A. Hecker, H. Labiod, G. Pujolle, H. Afifi, A. Serhrouchni and P. Urien, "*A New Control Access Solution for a Multi-Provider Wireless Environment*", ICTSM 10, 2002, Monterey, CA, USA.

[4] B. Zouari, H. Afifi, A. Hecker, H. Labiod, G. Pujolle, P. Urien, "*A Novel Authentication Model Based on Secured IP Smart Cards*", IEEE ICC 2003, Anchorage, Alaska, USA.

[5] W. Laouiti, K. Cordoso, A. Hecker, M. Genet, B. Jouaber, H. Labiod, D. Zeghlache, "*RESACO: An Open and Programmable Multi-Domain Platform for Cooperative and Auto-Configurable Networks*", ICWN 2004, Las Vegas, Nevada, USA.

[6] A. Hecker, H. Labiod, "*An Efficient Micromobility Implementation for 802.1X WLANs*", IEEE PIMRC 2004, Barcelona, Spain.

[7] A. Hecker, H. Labiod, "*A New EAP-based Signaling Protocol for IEEE 802.11 Wireless LANs*", IEEE VTC Fall 2004, Los Angeles, CA, USA.

[8] A. Hecker, H. Labiod, "*Pre-authenticated Signaling in Wireless LANs using 802.1X Access Control*", IEEE GLOBECOM 2004, Dallas, TX, USA.

[9] A. Hecker, E.-O. Blass, H. Labiod, M. Zitterbart, "*COMPASS: Decentralized Management and Access Control for WLANs*", accepted at 10th IFIP Conference on Personal Wireless Communications (PWC), August 2005, Colmar, France.

National Refereed Conference

[10] G. LeGrand, A. Hecker, F. Springinsfeld, "*Architecture flexible de réseau sans fil WiFi sécurisé*", SAR 2004, La Londe, France.

[11] A. Hecker "*Network Access: Centralized Architecture for Secure Port Access to Wireless Networks*", SAR 2002, Marrakech, Morocco.

Technical Reports

[12] C. Chiollaz, G. LeGrand, A. Hecker, F. Springinsfeld, S. Naqvi, Y. Denef, « *Politique de sécurité pour une communauté ouverte dans un espace ouvert délimité* », Délivrable D3a RNRT InfRadio, INFRES, ENST, 2004.

Index

Appendices

Appendix A Security Issues in 802.11

A.1. Insecurity of the Integrated 802.11 Standard Measures

Among the existing WLAN standards, only products of the IEEE 802.11 norm are available on the market. IEEE 802.11 standard defines different security measures. The exact definitions are out of the scope of this review. We assume that the reader is familiar with the IEEE 802.11 standard.

To sum up briefly: the central 802.11 security mechanism is called WEP. It is supposed to achieve "a security level comparable to that of the wired world" [14]. WEP is based on the RC4 stream cipher [38] and used for both authentication and packet encryption. Additionally, each 802.11 network has an identifier (SSID). Its knowledge is necessary to connect to a WLAN configured in the so-called "closed network" mode.

Between 2000 and 2002, different articles have been published referring to the 802.11 security resulting in every security definition of the IEEE standard being broken.

Attacks on WEP and other security measures

One of the first publications on the WEP security first appeared as an internal IEEE document. In [152], the author explains that the main deficiency of the WEP standard is not related to its default key length of 40 bits. The author demonstrates problems with the usage of the initialization vector (IV) within the WEP standard. He introduces known-plaintext and chosen plaintext attacks [38], which rely on this improper IV usage.

Consequently, these attacks are independent of the key length. The attacks are then extended to a cipher text only attack [38] by describing methods to build dictionaries containing <*IV, pseudorandom*> pairs. The analysis done in the paper shows that the usage of a stream-cipher is a wrong choice for a unreliable datagram service: the RC4 brick has to be re-initialized with every datagram for synchronization reasons (packet loss immediately results in de-synchronization of sender and receiver). The author insists on the argumentation against the RC4 describing a known issue with weak keys appearing every 256 keys. The paper questions whether the base key can be found in this particular setup which would result in complete collapse of the security architecture. The author suggests using some modern symmetric cipher instead of RC4 and proposes AES in OCB mode [38] for the WEP successor.

Then, in [153] a Berkeley team analyzes the WEP problems and shows that the dictionary necessary for the attack described in [152] has to be about 24GB. The authors describe the problems related to the key stream reuse and illustrate the implications of the Birthday Paradox on the IV-usage in WEP (leading to an IV collision after about 5000 packets). The lack of key management is pointed out as a main deficiency. The authors present a possibility to modify the content of WEP-encrypted messages without key knowledge. Thus, if the content is known, there is a possibility to change the packet to a correctly encrypted packet carrying attacker-provided data. Message injection without key knowledge is described after that. Using this possibility, the authors show that the WEP based authentication can be completely bypassed by any attacker having eavesdropped and recorded some previous authentication exchange. The paper explains how that blindly recorded authentication exchange of the challenge/response pair can be re-used indefinitely to gain network access – without any key knowledge. The authors then develop a method to decrypt WEP messages without key knowledge. They eventually conclude that WEP has been developed without any participation of the cryptographic community and that its authors did not have enough understanding of cryptographic primitives.

Similarly, in [154] the authors describe all 802.11 access control mechanisms and demonstrate experimental results proving that all these measures are not sufficient to provide a reasonable level of security. They also criticize the lack of key management and independently find the authentication attack mentioned in [153].

Attacks on RC4

In 2001, the cryptographic community presents a cryptanalysis of the key scheduling mechanism of the RC4 [155]. The authors present two attacks on the RC4 key. The first is based on the so-called invariance weakness. The second is based on the known-IV weakness. Both mechanisms are defined in [155]. Both attacks reconstruct a randomly chosen RC4 key with a complexity significantly lower than that of a brute force attack. The introduced known-IV attack is even independent of the key length. In the appendix, the authors explicitly mention the theoretical applicability of the second attack to the WEP-like cryptosystems, making it possible to find the base key and thus confirming the apprehension mentioned in [152].

Finally in [156], an AT&T team describes their practical implementation of a passive WEP-attack based on [155], i.e. the recovery of the secret key by a passive traffic observation. The authors describe their experiences and state that the whole work related to the attack (i.e. test-bed installation, chipset debugging and implementation) took them less than a week. Experience shows that an actual attack on a 40bit key takes about 2 hours and scales linearly with the WEP-key bit length (about 5-6 hours for 108bit keys presuming full traffic on a 802.11b WLAN).

Tools

Today, multiple different WEP crack tools are available for a free download on the Internet. These tools include e.g. AirSnort (http://airsnort.shmoo.com) and WEPCrack (http://wepcrack.sourceforge.net). The AirJack library (available for download at http://sourcefourge.net/projects/airjack) permits raw frame injection and reception and can be used to setup rogue APs, man-in-the-middle attacks, diverse DoS attacks, etc.

A.2. New Security Paradigms for 802.11

The security problems in 802.11 are not limited to the existence of the flaws. Even if there had been no flaws, the integrated security mechanisms would not have been sufficient. The problem is in the concept presuming only two user groups: the users who know the network-wide secret and the users who do not. This is not fine-granular enough for the broad spectrum of the planned usage scenarios for the modern WLANs.

What is needed, is a security definition which allows the representation of a defined network-wide security policy on the 802.11 level. This translates to a personalized security, i.e. per-user security, where all security functions and their levels are determined by the user profile. That is however not straightforward to integrate logically, since 802.11 entities are not users but network adapters.

For that reason, all recent security approaches such as Wi-Fi Alliance's WPA and the expected IEEE 802.11i draft supplement to the base 802.11 standard [81], use a port-based access control [80]. Port-based access control blocks all traffic on a (logical) port until some condition is true. The condition for the port opening is a successful user authentication over methods defined in IEEE 802.1X [80].

IEEE 802.1X defines the EAPOL which specifies a direct transport of IETF's EAP frames within 802 MAC frames. In the WLAN case, the access points act as authenticators. The authenticators are designed to open and close the port but can not authenticate users. The authentication is carried out by another logical entity called authentication server (AS). This is possible by blindly copying the content of the transported EAP frames sent by the supplicant (user, station, etc.) to the AS. The communications between the AS and the authenticators is usually based on IETF's AAA protocols like e.g. RADIUS. AS is designed to authorize and authenticate users according to their profiles but does not implement any link-relevant mechanisms. To enable link security between the authenticator and the supplicant, AS adds key material to its connection accept reply in case of successful verification. The access point uses this key material to derive the necessary 802.11 keys. Supplicants proceed adequately.

With this design, the link security mechanisms (encryption, integrity, etc.) can be setup dynamically, per-session. These are bound to the previous authentication and thus fulfill the typical modern security requirements. Using the available key material, the link keys can be changed periodically during the session (rapid re-keying). These considerations and the necessary mechanisms are the main part of the newer standardizations like e.g. IEEE 802.11i. These measures alone resolve the most urgent problems with the integrated 802.11 security and WEP protocol. To provide a more reliable link security, both WPA and 802.11i additionally redefine 802.11 mechanisms providing new protocols such as TKIP, CCMP, etc.

Appendix B AUTHENTIS Implementation

In our implementation we used the following components:

- PC-Server running under GNU/Debian Linux, Debian 2.2, Linux kernel 2.2.19.
- Linux FreeS/WAN: IPSec and IKE implementation (http://www.freeswan.org). A stable and easy-to-configure IPSec stack for Linux.
- FreeRADIUS server (http://www.freeradius.org): This is a GPL RADIUS implementation featuring the support for the most current Unix-derivates. It supports diverse RADIUS attributes, proxying (roaming support), a lot of authentication methods, EAP among others, SQL and LDAP.
- Cisco 340 Aironet: This access point in the Cisco's program supports both 802.1X and RADIUS.

All changes which we have applied are limited to the AAA servers.

Appendix C EAP-SIG Implementation

C.1. General

To implement EAP-SIG, no changes are required to the deployed numerous access points. However, an EAP client and an EAP server implementations are required. We made changes to the existing implementations. We namely used the following components.

Client: implements both signaling (EAP-SIG) and authentication part (802.1X supplicant)

- IBM ThinkPad Laptop with a Cisco Aironet 350
- Linux OS with kernel 2.4.26
- Xsupplicant v1.0pre2

Server: implements both SIGS and 802.1X AS

- The host "test" from the INFRADIO project
- Linux OS with kernel 2.4.27
- AAA Server FreeRADIUS version 1.0.1

We used a virtual AP on the basis of the Cisco AP1200 deployed within the the operational INFRADIO platform. The client part uses an implementation with two main modes based on a finite state machine.

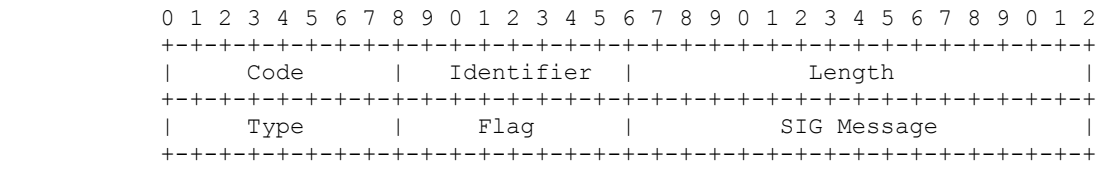
C.2. Packet Format

EAP-Response/Identity

The response to the EAP-Request Identity message issued by the Access Point as part of its 802.1X implementation, in our current implementation the client uses the extension of the EAP Response Identity message. The field containing the identity string is extended by a zero byte and a following EAP-SIG payload. In the following, these packets are denoted as EAP-Resp-Id ("*[string]*").

EAP-SIG

The basic format of the EAP-SIG packets when used as a distinct EAP method is presented in the following:



`Code`

- 1 - Request
- 2 - Response

`Identifier`

- Identical to any usual EAP method

`Length`

- Length(Code+Identifier+Length+Type+Flag+SIG Message)

`Type`

- 10 EAP-SIG

`Flag`

- Used for packet fragmentation. Not used in the current implementation and thus always contains the hexadecimal value 0x1.

`SIG Message`

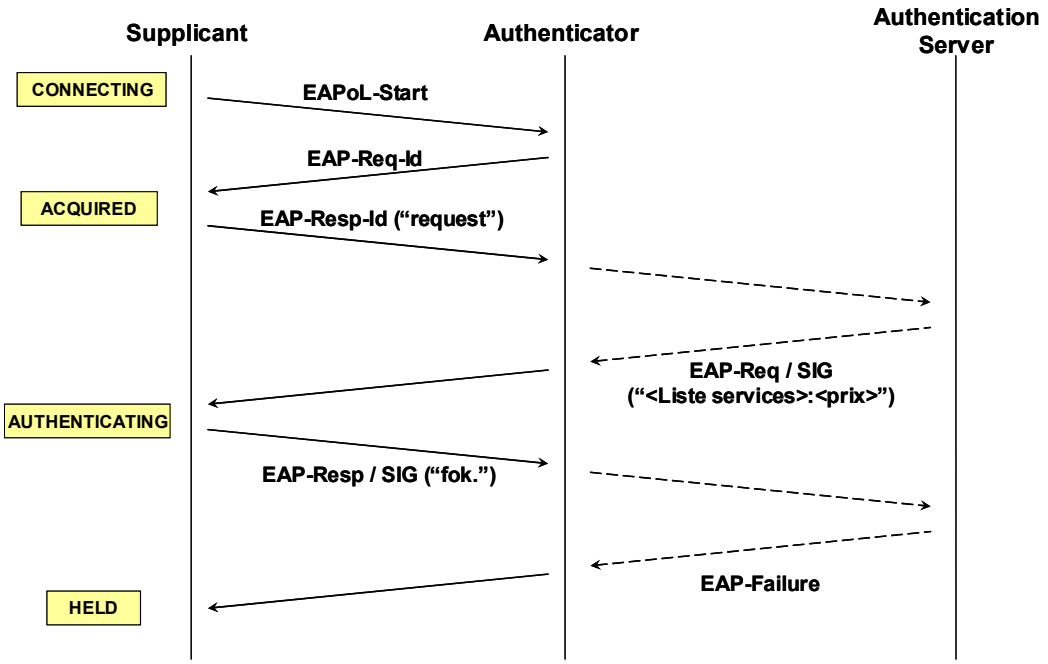
- This field contains the EAP-SIG data exchanged between the supplicant and the authentication server. The scenarios presented below illustrate the currently used protocol.

In the following these are denoted as EAP-Req / SIG (“[SIG Message]”) and EAP-Resp / SIG (“[SIG Message]”) respectively.

C.3. Scenarios

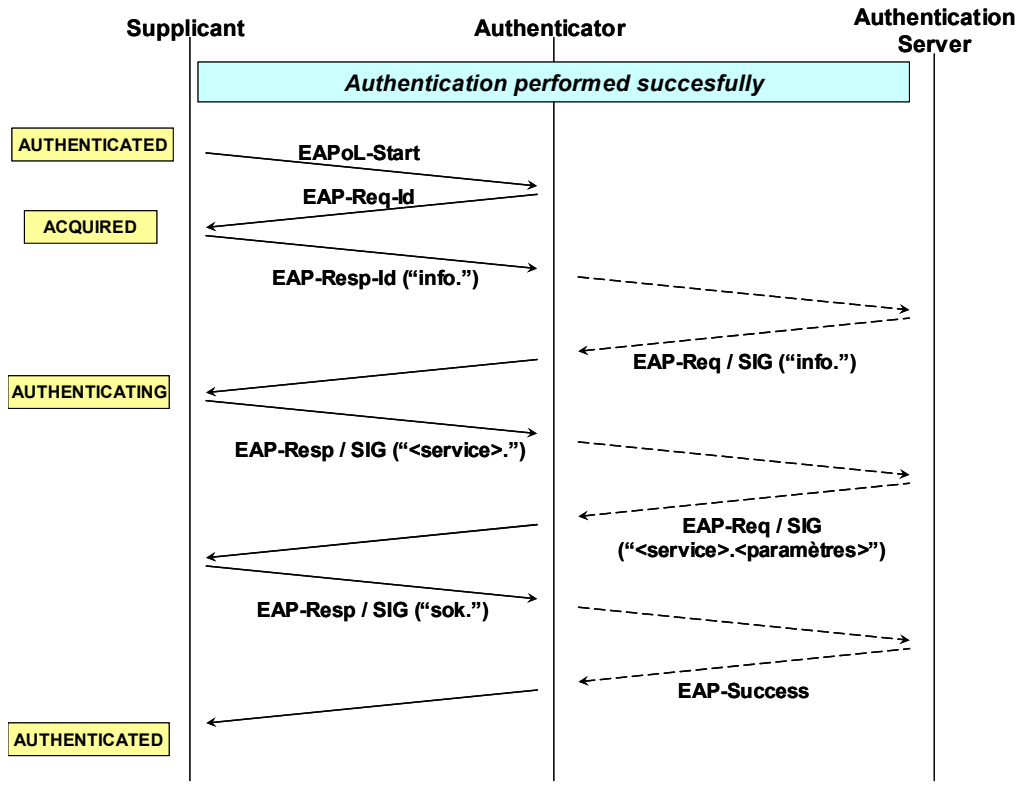
Mode 1

A user wants to discover services to which she has the authorization to access and the associated prices. The Figure below represents the message exchanges in mode 1 i.e. the early signaling (or pre-authenticated signaling). The respective state of the state machine of the client is shown at the left.



Mode 2

The Figure below represents the message exchanges in mode 2 i.e. after the user has connected to a chosen network (in-session signaling).



The last exchange (EAP-Resp / SIG ("sok.") → EAP-Success) is a security flaw since our demonstration implementation currently does not use the full potential of the cryptography. For simplicity reasons, we use EAP-MD5 as EAP authentication method which does not establish any cryptographic material between the client and the server. Thus, the success message would be produced by EAP-SIG without checking the authentication state. To close this security breach, we bind the authentication process to an expiration clause and add the last authenticationt timepoint to a special file used by a slightly adapted EAP-MD5 module. The latter writes the absolute time of the last successful authentication in the file `/usr/local/etc/raddb/sig_security`. This entry is of the following form:

```
[user identity] [authentication time]
```

When the EAP-Resp-Id ("info.") packet arrives at the server, the EAP-SIG module checks this file and verifies that the difference between the current time and the stored time is less than the defined Session Timeout. We use a global Session Timeout setting of 5 minutes for all users. Only if this condition is fulfilled, the EAP-SIG module replies with a success message and the demanded service description.

C.4. EAP-SIG User Manual

Server-side: FreeRADIUS

EAP-SIG is implemented as a module under FreeRADIUS. It has to be referenced in the central configuration file of FreeRADIUS, which is per default the following:

```
/usr/local/etc/raddb/radiusd.conf
```

Our module uses two additional files in the same directory `/usr/local/etc/raddb`:

- `customers`: specifies the services and the prices for each user. The used syntax is line based with the lines of the form <*username [servicelist] [pricelist]*>. Example line:
  `artur [smtp,http,ssh] [0,6]`
- `services`: contains service description information necessary to configure the services in the user equipment. The syntax follows lines of the form <*servicename [service description]*>. An example line is:
  `smtp [137.194.47.84:2000,user,passwd]`

User-side: XSupplicant

1) Configuration

Before starting the client, it has to be properly configured to use the EAP-SIG method. Currently, EAP-SIG has to be before other EAP methods in the configuration file. In the following we show an example configuration file using EAP-SIG and EAP-MD5 for demonstration purposes.

The value of the identity field is transmitted to the signaling and authentication server and used by the latter server as user identity.

```
default
{
        allow_types = all
        ssid = <BEGIN_SSID>ENST-INFRES-Test<END_SSID>
        identity = <BEGIN_ID>artur<END_ID>
        eap-sig {
                }
        eap-md5 {
                username = <BEGIN_UNAME>artur<END_UNAME>
                password = <BEGIN_PASS>F5gl8Yip<END_PASS>
                }
}
```

2) Starting the extended XSupplicant under Linux OS in mode 1

```
> ifconfig [iface] down
> iwconfig [iface] essid "[SSID]" enc [on | off]
> ifconfig [iface] up
> xsupplicant -c [config file] -i [iface] -s 1
```

The last parameter (`-s`) followed by `1` activates the datagram mode.

3) Starting XSupplicant under Linux OS in mode 2

```
> ifconfig [iface] down
> iwconfig [iface] essid "[SSID]" enc [on | off]
> ifconfig [iface] up
> xsupplicant -c [config file] -i [iface] -s 1 -m [service]
```

The parameter `-s` followed by `2` activates the dialog mode and discovers service information on service identifier following the `-m` parameter.

C.5. Code description

FreeRADIUS

All modified files and additions are under `./src/modules/rlm_eap/` in the FreeRADIUS source package. We have added the rlm_eap_sig module under `types/rlm_eap_sig`.

Modifications:

- `libeap/eap_types.h` and `libeap/eapcommon.c`: integration of EAP-SIG in the module `rlm_eap` (`Type = 10`)
- `eap.c`, in function `eaptype_select`: in case when the module receives the `EAP-Resp-Id` packet, if there are data appended to the identity string (e.g. strings "request" or "info."), the variable `default_eap_type` is set to the type value of EAP-SIG.

- `types/rlm_eap_md5/rlm_eap_md5.c`, in function `md5_authenticate`: when the EAP-MD5 authentication succeds the <*user, current system time*> tuple is written to the file `sig_security`.

XSupplicant

Modifications par rapport au code de base

- `eap.c`: EAP-SIG message construction
- `statemachine.c`: finite state machine
- `xsup_driver.c`: command line parameters

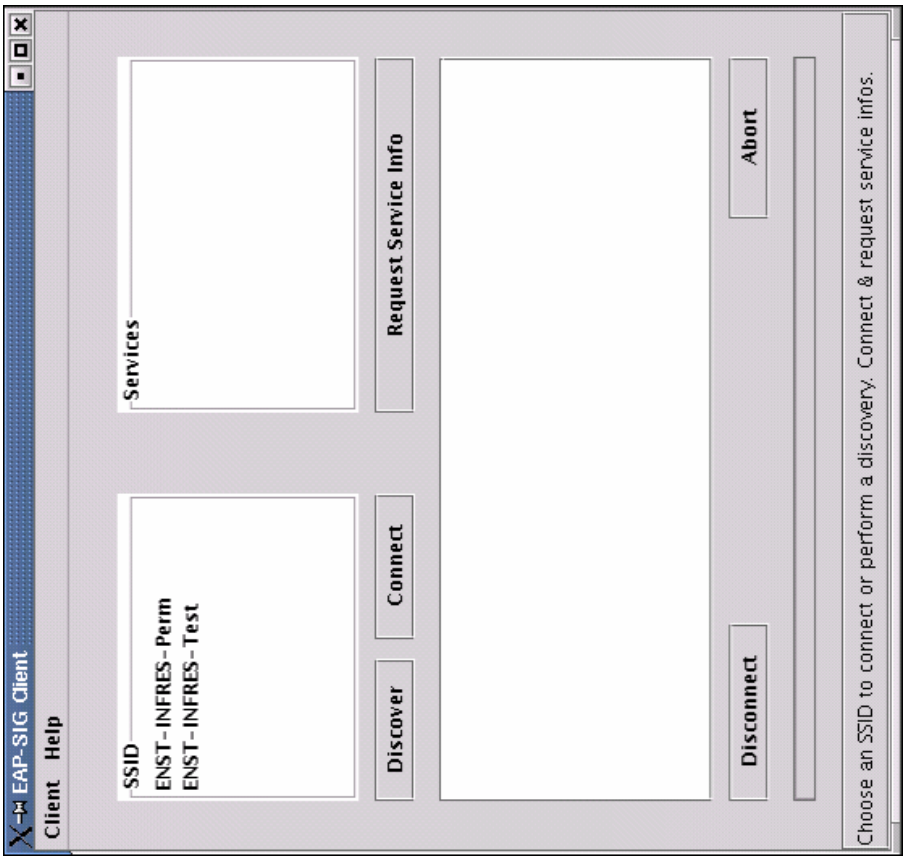
Other files like `config.c`, `profile.h`, etc. are slightly modified (variable definitions, data structures, etc.

C.6. Graphical User Interface

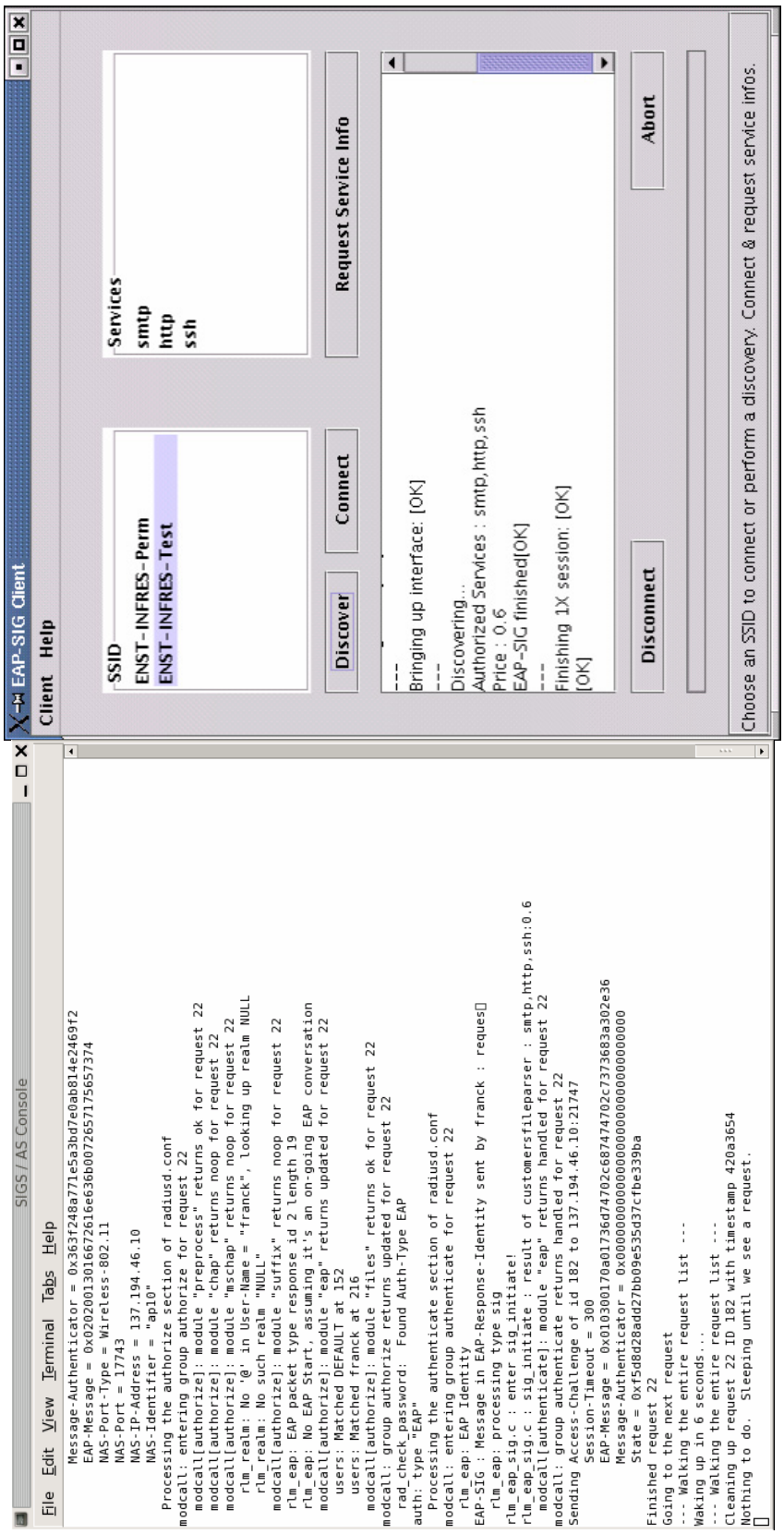
To simplify the user part of our SIG-supplicant, we developed a graphical user interface (GUI) in Java. The GUI launches the SIG-supplicant in the necessary modes and automatically extracts the information from the incoming packets. It thus simplifies network discovery, connection and in-session service discovery.

In the following we present different scenarios with the screenshots of the developed GUI and the correspondent screenshots of the SIGS console.

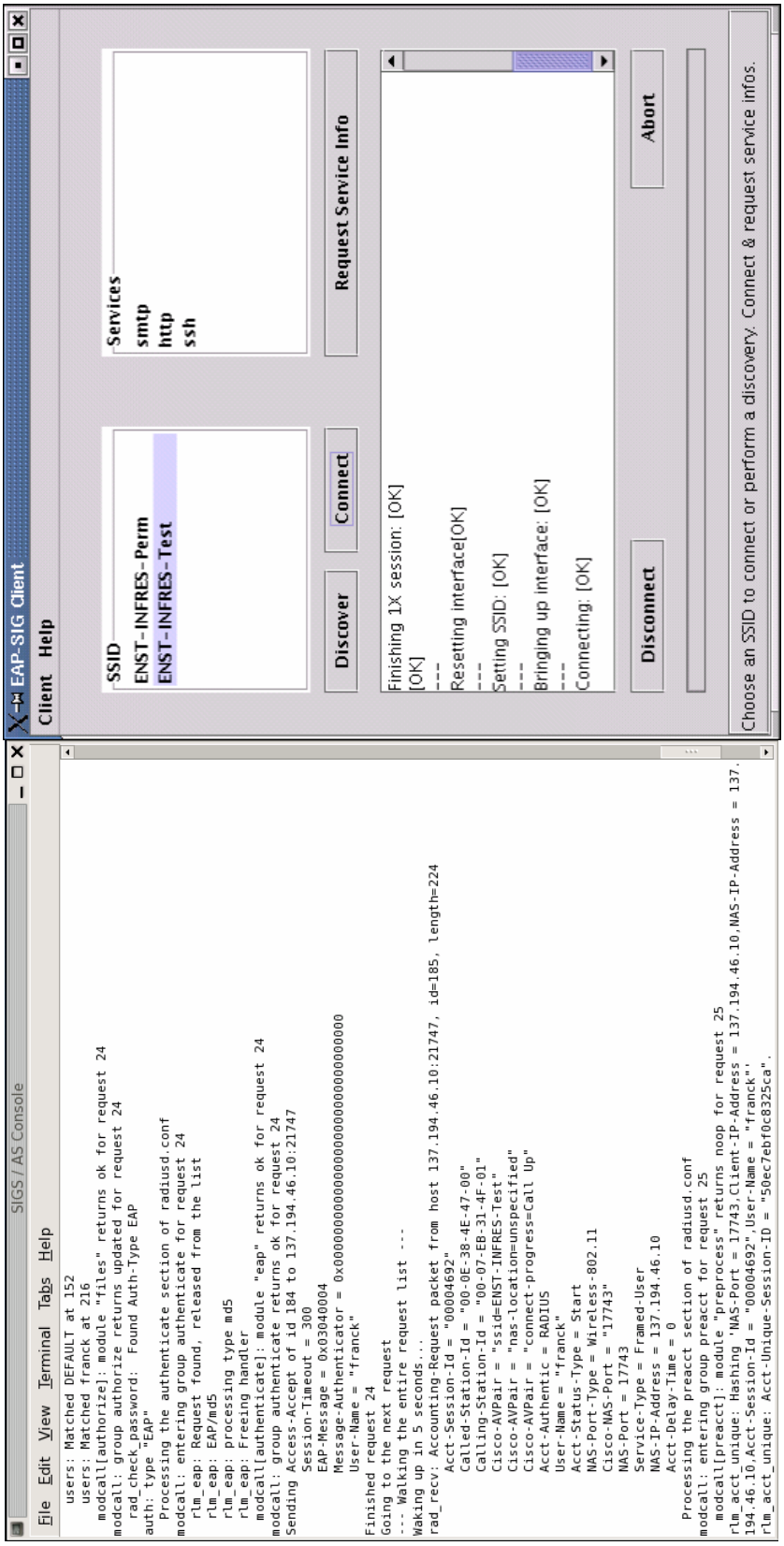
Scenario 1: EAP-SIG Start



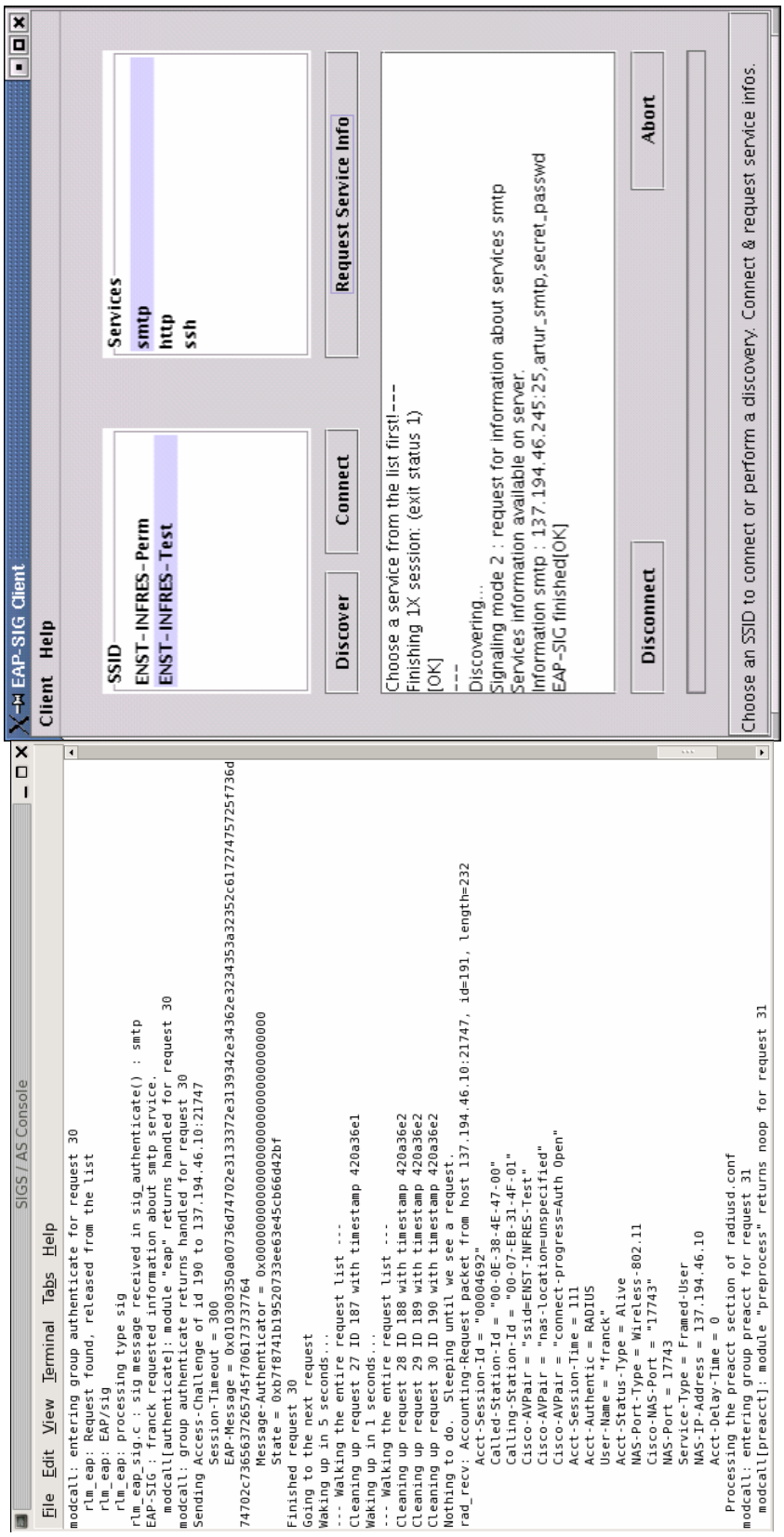
Scenario 2: Early Service Discovery



Scenario 3: Network Connect



Scenario 4: In-Session Service Property Discovery



C.7. Bugs and ToDo

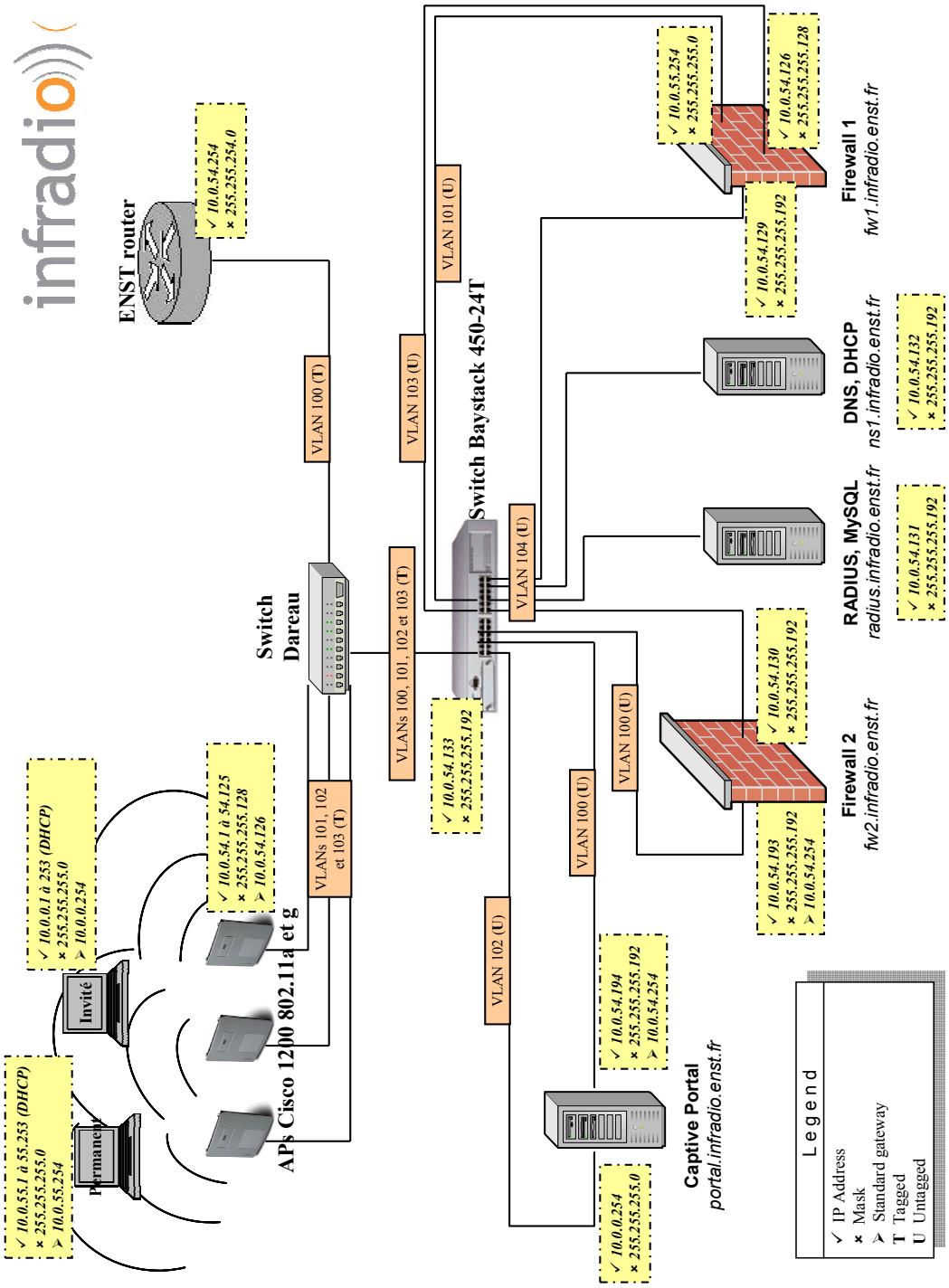
- Add fragmentation
- Add real cryptography based on the user-server trust
- In mode 2, when the client requests service information the server currently does not check if this particular user has the authorization for this service (profile based discovery is not yet implemented).

Appendix D INFRADIO Implementation

The INFRADIO platform is described in all details in [121] and [122]. We principally (but not exclusively) use the open source software. The deployed access points are Cisco Aironet 1200 APs. The switches provide the necessary VLAN switching, bringing the tagged packets to our platform. The lattter is separated in two main parts.

The public part includes a single machine installation (Linux OS with kernel 2.4.27) with an extended captive portal. The captive portal features source network address translation (n:1 SNAT), dynamic filtering and automatic WWW request redirect to an integrated Web server. We use the NoCat package (http://www.nocat.net) as the base implementation for our captive portal. The SNAT feature and the dynamic packet filtering are provided by the Linux kernel. The Web server is an Apache open source Web server (http://www.apache.org). All users are automatically redirected to the internal pages explaining access conditions to our public 802.11 service. An optional authentication (provided over NoCat) enables further features such as e.g. a VPN service (PPTP based, http://www.poptop.org). Following the delegation concept, the authorized users can create accounts of temporary validity when connecting to the internal interface.

The private part consists of a demilitarized zone (DMZ) protected by two Linux based firewalls. The DMZ contains the business, control and network plane entities. The network plane is implemented by the firewalls acting as routers. These are principally configured to only allow connections to the DMZ (but not DMZ traversal, except for the some authorized users as required according to our security policy). The control plane consists of a DNS server (authoritative for the subdomain `infradio.enst.fr`), a DHCP server for the private 802.11 service users, a RADIUS server (http://www.freeradius.org) with an associated user database (http://www.mysql.org) and a backup solution. The business plane consists of a commercial software for database and RADIUS server management (http://www.wave-storm.com). Traffic observation is carried out by another commercial software (http://www.qosmos.com). The complete current platform architecture is shown below.


infradio
ENST router
✓ 10.0.54.254
× 255.255.254.0
Switch Dareau
VLANs 101, 102 et 103 (T)
VLAN 100 (T)
VLANs 100, 101, 102 et 103 (T)
VLAN 101 (U)
VLAN 103 (U)
Switch Baystack 450-24T
VLAN 104 (U)
VLAN 100 (U)
VLAN 102 (U)
APs Cisco 1200 802.11a et g
✓ 10.0.0.1 à 253 (DHCP)
× 255.255.255.0
➢ 10.0.0.254
✓ 10.0.54.1 à 54.125
× 255.255.255.128
➢ 10.0.54.126
✓ 10.0.55.1 à 55.253 (DHCP)
× 255.255.255.0
➢ 10.0.55.254
Invité
Permanent
✓ 10.0.54.133
× 255.255.255.192
✓ 10.0.54.194
× 255.255.255.192
➢ 10.0.54.254
✓ 10.0.0.254
× 255.255.255.0
Captive Portal
portal.infradio.enst.fr
✓ 10.0.54.130
× 255.255.255.192
✓ 10.0.54.193
× 255.255.255.192
➢ 10.0.54.254
Firewall 2
fw2.infradio.enst.fr
✓ 10.0.55.254
× 255.255.255.0
✓ 10.0.54.126
× 255.255.255.128
✓ 10.0.54.129
× 255.255.255.192
Firewall 1
fw1.infradio.enst.fr
RADIUS, MySQL
radius.infradio.enst.fr
✓ 10.0.54.131
× 255.255.255.192
DNS, DHCP
ns1.infradio.enst.fr
✓ 10.0.54.132
× 255.255.255.192
Legend
✓ IP Address
× Mask
➢ Standard gateway
T Tagged
U Untagged

Appendix E RESACO Implementation

In our implementation we used the following components:

Access Point:

- PC running Linux OS kernel 2.4 with a 802.11 network adapter
  - Linux kernel network filter (iptables, http://www.netfilter.org)
- HostAP project (driver, http://hostap.epitest.fi)
- Intel's UPnP SDK (http://www.intel.com/technology/UPnP/index.htm)

Local RADIUS Server:

- FreeRADIUS server (http://www.freeradius.org).
- PHP (http://www.php.net)
- Apache Web server (http://www.apache.org)

User services:

- Squid Web Proxy (http://www.squid-cache.org)
- VideoLAN (http://www.videolan.org)